互联网金融系列丛书



大数据金融与征信

何平平　车云月　编　著

大数据+金融落地实践之作
汇集大数据金融专家思想精华

大数据时代，深度剖析当下及未来征信行业发展机遇与路径

清华大学出版社

互联网金融系列丛书

大数据金融与征信

何平平　车云月　编著

清华大学出版社
北　京

内 容 简 介

本书面向金融应用，系统地阐述了大数据金融与征信本身及其在现实生活中的应用，具有全面性、实用性和前瞻性等特色。全书共 8 章，第 1 章和第 2 章阐述大数据金融及大数据技术相关的基础知识问题，是后面章节的基础。第 3 章至第 6 章详细介绍大数据在银行业、证券业、保险业及互联网金融行业中的应用，是本书的主要内容。第 7 章重点阐述大数据在征信中的实际应用，是本书的另一重点问题，也是当代大数据研究的热点问题。第 8 章特别强调中国金融信息安全，这是大数据金融与征信的发展进程中不可避免的问题。本书力争把大数据与其实际应用糅合在一起介绍，力求活学活用。

本书可以作为高等学校互联网金融院系课程教材，也可供互联网金融研究者、从业者、管理人员参考所用。

图书在版编目(CIP)数据

大数据金融与征信/何平平，车云月编著. —北京：清华大学出版社，2017
(互联网金融系列丛书)
ISBN 978-7-302-48440-0

Ⅰ. ①大… Ⅱ. ①何… ②车… Ⅲ. ①金融—数据管理—研究 ②互联网络—应用—金融—研究 Ⅳ. ①F830.49

中国版本图书馆 CIP 数据核字(2017)第 225246 号

责任编辑：杨作梅
封面设计：杨玉兰
责任校对：王明明
责任印制：沈　露
出版发行：清华大学出版社
　　网　　址：http://www.tup.com.cn, http://www.wqbook.com
　　地　　址：北京清华大学学研大厦 A 座　　**邮　　编：**100084
　　社 总 机：010-62770175　　**邮　　购：**010-62786544
　　投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
　　质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市铭诚印务有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张：**18.25　　**字　数：**440 千字
版　　次：2017 年 10 月第 1 版　　**印　次：**2017 年 10 月第 1 次印刷
印　　数：1~3000
定　　价：42.00 元

产品编号：076698-01

前言

大数据金融是大数据在金融领域的重要应用。大数据金融市场前景广阔，预计未来 5 年到 10 年，金融大数据产业将迎来黄金增长期，大数据也将成为助推“大众创业、万众创新”浪潮的有力抓手。

本书为适应高等学校互联网金融专业人才培养的需要，从理论联系实际的原则出发，以大数据的实际运用为导向，对大数据在金融各行业的应用做了全面系统的介绍。

全书共分为 8 章，包括大数据金融概述、大数据相关技术、大数据在商业银行中的应用、大数据在证券行业中的应用、大数据在保险行业中的应用、大数据在互联网金融中的应用、大数据征信、大数据与中国金融信息安全。

由于大数据金融刚刚兴起，可供参考的资料不多，本书也仅仅是在这方面的一个探索，故全书整体框架以编者自己的思路进行呈现。本书以应用特别是金融领域前沿的应用为导向，以在各行业的实践为主线展开。本书内容新颖全面，论述问题极具现实意义。本书可以作为高等院校互联网金融专业相关课程的教材，也可供互联网金融研究者、从业者、管理人员参考。

全书主要有以下两大特点。

(1) 内容全面。

本书以大数据为出发点，结合国内外的发展现状及最新模式，系统地介绍了大数据在银行业、证券业、保险业、互联网金融行业及征信中的应用，并强调了在应用过程中，中国金融信息安全的重要性及保障机制。本书内容涵盖面极广，有效地为各行各业的读者提供了大数据金融与征信的宏观视图。

(2) 体例新颖。

本书秉承着注重实际运用的宗旨，编写体例上彰显了可读性和互动性。每章前有“本章目标”和“本章简介”，每章末有“本章总结”和“本章作业”。书中除了理论教学，还配有相关案例和解析，使理论与实践相结合，通俗易懂，开拓了学生的视野，可以更好地满足培养既懂专业知识又能运用所学知识解决实际问题的“复合型”经济人才需求。

本书由新迈尔(北京)特技有限公司组织研发，由何平平拟定大纲并进行统稿，湖南大学互联网金融研究所组织撰写。本书由何平平、车云月担任主编，以下研究生也参与了本书的编写：王杨毅彬、周春亚、张童、刘诗雨、刘晶宇。

本书编写过程中参考了大量的文献资料，有些已经在书后的参考文献中标注，而有些没有，在此一并表示感谢。囿于时间和个人能力，书中难免有疏漏和不妥之处，敬请读者批评指正。

何平平

《互联网金融系列丛书》编审委员会

目录

目录

第 1 章

大数据金融概述

本章目标

- 掌握大数据的内涵与特征
- 了解大数据产生的背景
- 掌握大数据的类别
- 了解大数据的价值和应用领域
- 掌握大数据金融的内涵特点
- 掌握大数据金融相对于传统金融的优势
- 了解大数据给金融业带来的大变革
- 了解大数据给征信业带来的大变革
- 了解互联网大数据中的应用
- 掌握大数据金融的两种模式
- 了解大数据金融信息安全

本章简介

随着计算机技术和互联网的发展，大量的音频、图片、视频等结构化数据和半结构化数据不断涌现，传统的数据处理技术已经难以应对，因此大数据的概念应运而生。随着大数据技术的成熟，大数据已经广泛应用于商业、通信、医疗、金融等领域，给各行各业带来了巨大的价值。

近几年，大数据浪潮迅速席卷全球，数据成为企业重要的生产要素和战略资产，拥有大数据资产的企业将在竞争中占有优势。金融业本身就是基于数据与信息的产业，作为现代经济的核心，敏锐的金融行业正在积极拥抱大数据技术。大数据金融相对于传统金融有着无可比拟的优势，引起了金融行业广泛而深远的变革，包括银行业、保险业、证券业、征信业及互联网金融。

本章重点讲解大数据的内涵与特征、大数据的分类、大数据的处理流程以及大数据的价值和应用领域、大数据金融的内涵特点、大数据金融相对于传统金融的优势、大数据带来金融业和征信业大变革、互联网大数据的应用和大数据金融的两种模式。

1.1 大数据概述

在互联网中，大数据无处不在。无论是漫无目的的浏览网页、观看视频，还是发微博、聊微信，以及有目的性的搜索，基于每个用户都会产生数据，这些分散的数据汇集到网络中形成数据流，并最终聚集到网络服务提供商，形成大数据。

1.1.1 大数据的内涵与特征

1. 大数据与小数据

大数据(big data)是指在一定时间范围内无法用传统数据库软件进行采集、存储、管理和分析的数据集或数据群，需要通过新的处理模式才能体现出的具有高效率、高价值、海量、多样化特点的信息资产。利用数据挖掘分析技术可以使这些结构化、半结构化、非结构化的海量数据产生巨大的商业价值。小数据(small data)，或称个体资料，是以个体为中心，需要新的应用方式才能体现出的具有高价值、个体、高效率、个性化特点的信息资产。大数据和小数据有着本质的区别，虽然两者都是以创造数据价值为目的，但是在收集目的、数据结构、生命周期、分析方法及分析重点 5 个方面都存在着不同的定位。

1) 收集目的

小数据的目的性很强，往往是为了一个目标，制定规划进行收集、整理和分析，不会收集与其研究目的无关的数据。而大数据收集没有明确的目标，收集的数据范围更广，在数据采集阶段并不明确知道会产生什么结果。

2) 数据结构

小数据的数据基本来自相同的行业和领域，数据种类单一，结构统一，并采取一种有序排列的结构化方式。而大数据的数据来自不同的行业和领域，数据种类复杂，数据标准和格式有所不同，非结构化的数据居多，无法进行统一排序。

3) 生命周期

小数据的生命周期比较短，几乎只有几年的时间，待相关问题解决或相关项目结束之后，小数据一般会被删除。而大数据的工作主要是进行预测。只有基于完整的历史数据才能对未来进行相对准确的预测。因此，大数据的生命周期相对较长，大部分会被永久保留。

4) 分析方法

小数据采用一般的统计方法对收集的所有数据进行分析；而大数据因其复杂性一般通过分布式的方式进行分析，采用训练、学习、聚合、归一化、转化、可视化等多种不同的方法分析。

5) 分析重点

小数据是以个体行为数据为对象，主要是对个体数据信息进行全方位的精确的挖掘分析，重点在于深度；而大数据是以某个群体行为数据为对象，主要是对大范围大规模的数据处理分析，重点在于广度。

小数据不涉及大量的、急速的数据，或是繁多的信息种类，也没有隐含与大数据有关

的复杂化信息，并常以微观角度解释小型对象。而大数据则立于宏观角度，致力于表述宏观现象。简言之，用大数据得到规律，用小数据匹配个人。

2. 大数据的内涵

大数据的概念较为抽象。大数据中的“数据”是指广义的数据，不仅包括传统的结构化数据(即可以用二维表格表述的数据)，还包括非传统的非结构化数据(如视频、音频等)，大数据中的“大”既形容数据量多，也形容数据产生和变化的速度非常快。大数据的内涵主要体现在数据类型、技术方法和分析应用 3 个方面。

1) 数据类型方面

大数据不仅包括传统的结构化和半结构化的交易数据，还包括巨量的非结构化数据和交互数据，它是包括交易和交互数据集在内的所有数据集，如社交网站上的数据、在线金融交易数据、公司记录、气象监测数据、卫星数据和其他监控、研究和开发数据。

2) 技术方法方面

核心是从各种各样类型的数据中快速获取有价值信息的技术及其集成，依据大数据的生命周期的不同阶段可以将大数据处理技术分为大数据存储、大数据挖掘和大数据分析 3 个方面。大数据存储包括直接外挂存储(DAS)、网络附加存储(NAS)、存储域网络(SAN)等存储方式。大数据挖掘主要采用的是分布式挖掘和云计算技术。

3) 分析应用方面

重点是采用大数据技术对特定的数据集合进行分析，及时获得有价值的信息。常用数理统计方法进行数据分析，如可视化的数据分析工具。在数据分析过程中不仅需要计算机进行自动化的分析，还需要人工进行数据的选择和参数的设定。

3. 大数据的特征

大数据具有 5 个特征：大体量(Volume)、多样性(Variety)、时效性(Velocity)、准确性(Veracity)、价值性(Value)，如图 1.1 所示。

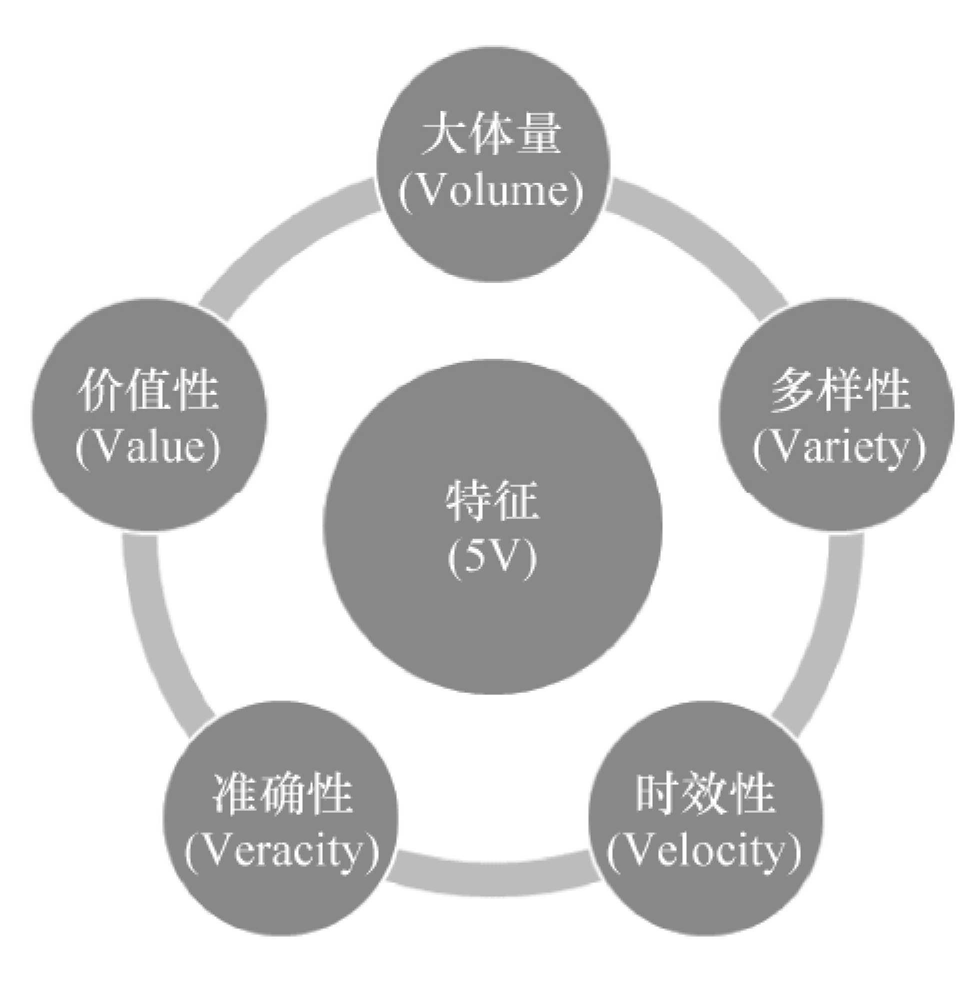


图 1.1 大数据的特征

1) 大体量

大体量，即数据量大，是大数据的基本属性。大数据一般是指 10 TB(1 TB=1024 GB)规模以上的数据量，甚至可从数百 TB 到数十数百 PB、EB 的规模。资料显示，百度首页导航每天需要提供的数据超过 1.5PB(1PB=1024TB)。导致数据规模剧增的原因有：①传感器等各种仪器获取数据的能力大幅提高，越来越多的事物特征可以被感知，这些特征数据将会以数据的形式被存储下来。②互联网的普及，使数据的分享和获取越来越容易，无论是用户有意还是无意的分享或浏览网页都会产生大量数据。③集成电路价格的降低，使很多数据被保存下来。国际数据资讯(IDC)公司监测，全球数据量大约每两年翻一番，预计到 2020 年，全球将拥有约 35ZB 的数据量(见图 1.2)，并且 85%以上的数据以非结构化或半结构化的形式存在。

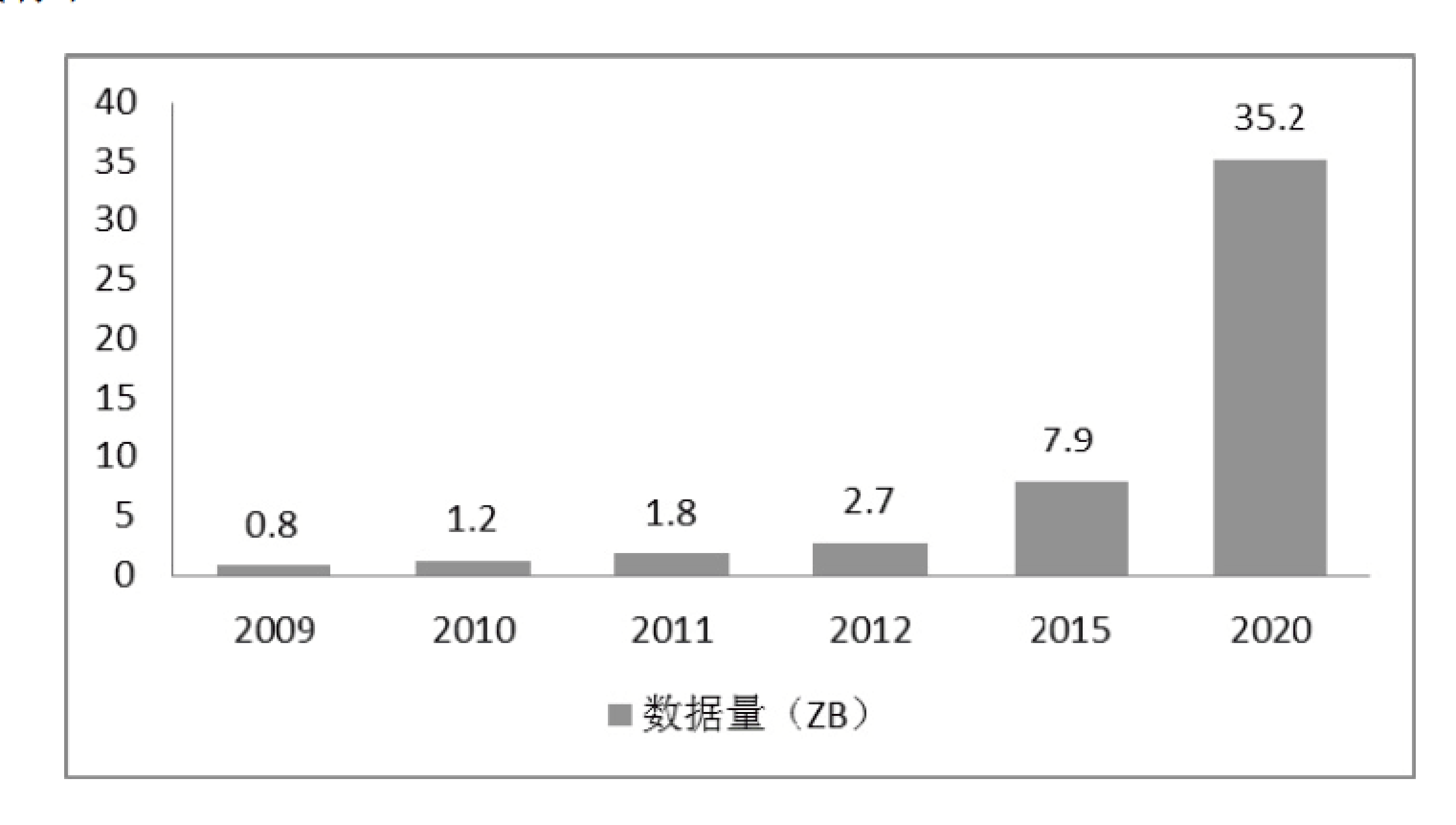


图 1.2 IDC 全球数据量使用情况及预测

2) 多样性

数据类型多样化是大数据的第二大特点。大数据包括各种格式和形态的数据。传统的数据大多是以二维表的形式存储在数据库中的文本类结构化数据。随着互联网的发展和传感器种类的增多，诸如网页、图片、音频、视频、微博类的未加工的半结构化和非结构化数据越来越多，以数量激增、类型繁多的非结构化数据为主。非结构化数据相对于结构化数据而言更加复杂，数据存储和处理的难度增大。目前，我国商业银行业务发展相关数据类型已从结构化数据扩展到非结构化数据。

3) 时效性

大数据的时效性是指在数据量特别大的情况下，能够在一定的时间和范围内得到及时处理，这是大数据区别于传统数据挖掘最显著的特征。大数据的流动速度快，当处理的数据从 PB 增加至 TB 时，超大规模的数据快速变化，使用传统的软件工具将难以处理。只有对大数据做到实时创建、实时存储、实时处理和实时分析，才能及时有效地获得高价值的信息。

4) 准确性

大数据的准确性是指保证处理的结果具有一定的准确性。结果的准确性涉及数据的可信度、偏差、噪声、异常等质量问题，原始数据的输入错误、缺失以及数据预处理系统的失效等会导致数据的不准确，进而分析得出一些错误的结论。因此，保证正确的数据格式

对大数据分析十分重要。

5) 价值性

大数据的价值性是指大数据包含很多深度的价值，对大数据的分析挖掘和利用将产生巨大的商业价值。数据量呈指数增长的同时，隐藏在海量数据中的有用信息却没有相应比例增长；相反，价值密度的高低常常与数据总量的大小成反比。这样反而使我们获取有用信息的难度加大。以商业银行监控视频为例，连续数小时的监控过程中可能有用的数据仅有几秒钟。

大数据的特征表明大数据不仅数据量巨大，种类繁多，对大数据的分析将更加复杂，更加追求速度，更注重时效性、准确性以及价值性。大数据不仅意味着数据总量的快速增长，其更大的意义在于：通过对大容量数据的交换、整合和分析，及时识别与发现新的知识，创造新的价值，带来“大知识”和“大发展”。作为一种重要的战略资产，大数据开启了一次全新的、重大的时代转型。

4. 大数据与传统数据的区别

大数据是以数量巨大、类型众多、结构复杂的数据集合以及基于云计算的数据处理和应用模式，通过数据的集成共享、交叉复用形成的智力资源和知识服务。大数据与传统数据在产生方式、存储方式、使用方式等方面都有所不同。

1) 产生方式

传统的数据是根据研究目的进行采集，采集的数据具有重要性。因为监管要求、业务逻辑或者技术便利，大数据具有“自产生”的特点，不需要特别的采集过程，比如搜索数据、交易数据等，尽管有些数据可能没有价值。

2) 存储方式

大数据的规模远远大于传统数据的规模。相对于传统数据库，量变引起质变，需要新的数据库技术来支持存储和访问。新型的大数据存储系统除了要具备高性能、高安全、高冗余等特征之外，还需具备虚拟化、模块化、弹性化、自动化等特征，才能满足具备大数据特征的应用需求。

3) 使用方式

传统数据是基于样本思维进行采集的，其分析方法主要是基于概率论理论和抽样理论。通常是通过这些样本数据推断总体，很难从这些数据中提炼出超出研究设计的知识。而大数据则是基于全体思维，所采集的数据基本能够代表整体，通过人工智能、神经网络等讲求高维和高效率的分析技术可以从这些详尽的数据中得出有价值的规律和知识。

5. 大数据的产生背景：计算机技术与互联网的发展

随着计算机的快速发展和互联网应用的成熟，数据量急剧增加，人类进入大数据时代。数据的采集、传输、存储、整合、管理、挖掘、分析等各项技术快速发展。

1) 计算机技术的发展

1946 年，第一台电子计算机的诞生开启了人类社会信息技术革命的序幕。截至目前，计算机技术的发展经历了大型主机、小型计算机、微型计算机、客户/ 服务器、互联网、云计算这六大阶段(见图 1.3)。

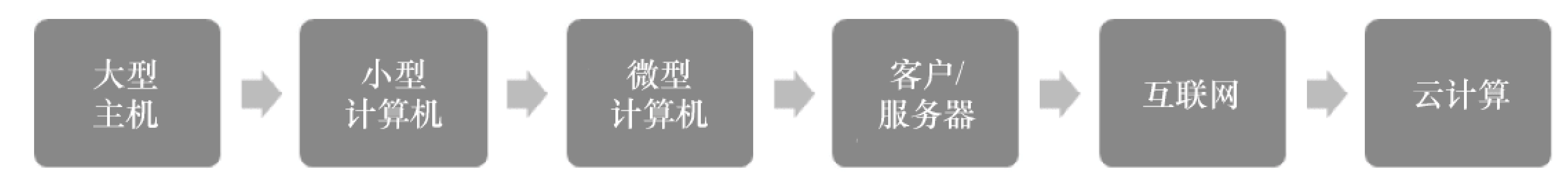


图 1.3　计算机技术经历的几个阶段

(1)　大型主机阶段(20 世纪 40—50 年代)。此阶段的计算机体型十分庞大，如第一台计算机由 18 800 个电子管组成，重量约 27 吨，占地约 150 平方米。在经历了电子管数字计算机、晶体管数字计算机、集成电路数字计算机和大规模集成电路数字计算机等发展历程后，计算机技术逐渐走向成熟。

(2)　小型计算机阶段(20 世纪 60—70 年代)。半导体和集成电路的改良使得大型主机经历了第一次缩小化，使用成本也因此降低，价格可被中小企业接受且能够满足中小企业的信息处理要求。现在很多企业使用的服务器都属于小型计算机，在体型上大于一般的个人计算机，小于大型主机。

(3)　微型计算机阶段(20 世纪 70—80 年代)。这个阶段是对小型计算机的缩小化，计算机已经缩小到可以放置在桌面上，因此被称为“微型计算机”或者“个人计算机”。1977 年美国苹果公司推出了 Apple 二代计算机，大获成功。1981 年 IBM 推出了 IBM – PC，经过不断的改良，功能不断加强，并占领了个人计算机市场，由此个人计算机得到了很大的普及。

(4)　客户机/服务器阶段。计算机的客户机/服务器结构起源于 20 世纪 60 年代，IBM 与美国公司建立了第一个全球联机订票系统，2000 多个订票终端被连在一起。在客户机/服务器结构中，网络的基础是客户机，核心是服务器，客户机通过服务器获得所需要的网络资源，其优点是能够充分发挥客户端的处理能力，减轻服务器的压力。

(5)　互联网阶段。1969 年，美国国防部研究计划署制定的协定将美国加利福尼亚大学洛杉矶分校、斯坦福大学研究学院、加利福尼亚大学和犹他州大学的 4 台主要的计算机连接起来，标志着计算机进入因特网阶段，即互联网阶段。此后，互联网经历了文本、图片、语音、视频阶段，带宽不断变快，功能越来越强大，这是人类迈向地球村坚实的一步。

(6)　云计算阶段。2008 年，“云计算”这个技术名词开始流行起来，它是一种基于互联网的计算方式，共享的软硬件资源和信息可以按照需求提供给计算机和其他设备。云计算阶段，计算机能力可以作为一种商品通过互联网进行流通。企业和个人不再需要购买昂贵的硬件，只需通过互联网来购买或者租赁计算能力，为所使用的计算功能付款。云计算囊括了开发、架构、负载平衡和商业模式等，是未来的软件业模式。

2)　互联网的发展

互联网不仅改变了传统的信息传播方式，也改变了人们的生活习惯。获取信息变得更加容易，足不出户便可了解世界新闻；沟通更加便捷，QQ、微信等网络工具将人们时刻联系在一起；购物消费更加容易，利用手机或电脑上网就可以快速实现商品交易。因此，互联网的发展不仅是一场信息革命，也是社会变革。根据第 38 次《中国互联网络发展状况统计报告》，截至 2016 年 6 月，中国网民规模达 7.10 亿人，其中手机网民规模达 6.56

亿人，占比 92.5%。网民行为因为互联网的发展更加多元化，文本、图片、音频、视频、地理位置等信息已经成为大数据增长最快的来源。

大数据与计算机技术和互联网的发展相辅相成。大体量的数据采集、存储、管理和挖掘因计算机和互联网技术的快速发展得以实现，数据的来源越来越丰富，形成信息流；大数据的信息流又通过社会生活和商业模式带动着资金流和物流的发展，进一步推动计算机与互联网技术的改进。大数据与计算机和互联网技术相互作用，相互促进，共同发展。

1.1.2　大数据的分类

大数据的种类很多，可以依照不同标准进行分类。

1. 按照大数据结构特征分类

按照大数据结构特征，可以将大数据分为结构化数据、非结构化数据和半结构化数据。

(1) 结构化数据。是指有结构的数据，也即行数据，在得到数据之前，其结构就是确定的。比如，传统的关系数据模型，可用二维结构表示。二维表中的数据就是典型的结构化数据，其结构事先通过数据模型的定义确定下来，在处理过程中不会改变。

(2) 非结构化数据。是指没有结构的数据，无法用数据库的二维逻辑结构来表现。包括所有格式的文档、文本、图片、视频、音频、各类报表以及标准通用标记语言下的子集 XML、HTML。它们通常没有数据模型，无法进行结构化处理。

(3) 半结构化数据。是指介于结构化数据和非结构化数据之间的数据。半结构化数据也是有结构的数据，与结构化数据不同的是，半结构化数据是先有数据，再有结构。半结构化数据一般是自描述的，数据的结构和内容混合在一起，没有明显的区分，其数据模型是数和图。常见的半结构化数据有 XML、HTML。

2. 按照大数据获取处理方式分类

按照大数据获取处理方式，可以将大数据分为批处理数据和流式计算数据。数据的批处理是指对数据进行批量的处理，如对数据进行成批的增加、修改、删除等操作。流式计算是指可以在实时处理的应用环境中，对大规模流动数据在不断变化的前提下进行持续计算、分析并能捕捉到有价值信息的分布式计算模式。流式数据具有实时性、易失性、突发性、无序性和无限性的特点。大数据的批处理和流式计算的区别如下表所示。

大数据批处理与流式计算的比较

性能指标	大数据流式计算	大数据批处理
计算方式	实时	批量
常驻空间	内存	硬盘
时效性	短	长
有序性	无	有
数据量	无限	有限
数据速率	突发	稳定

续表

性能指标	大数据流式计算	大数据批处理
是否可重现	难	易
数据精确度	较低	较高

3. 按照其他方式分类

按照大数据处理响应性能，可以将大数据分为实时数据、非实时数据和准实时数据；按照大数据关系，可以将大数据分为简单关系数据和复杂关系数据，如 Web 日志是简单关系数据，社会网络等具有复杂关系的图计算属于复杂关系数据。

1.1.3 大数据的价值

大数据最大的价值，是能够通过挖掘数据之间的相关性，把模糊的、隐含的、时滞性的问题，以可视化的、明确的、预演的方式展现出来，以便于决策和管理单元采取措施，改变所暴露的问题。这和传统的数据分析有着明显的不同。以往的数据分析或商业智能，更多的是面向过去已经发生的，而大数据是面向未来即将发生的。对金融行业来说，大数据主要有如下几点价值(见图 1.4)。

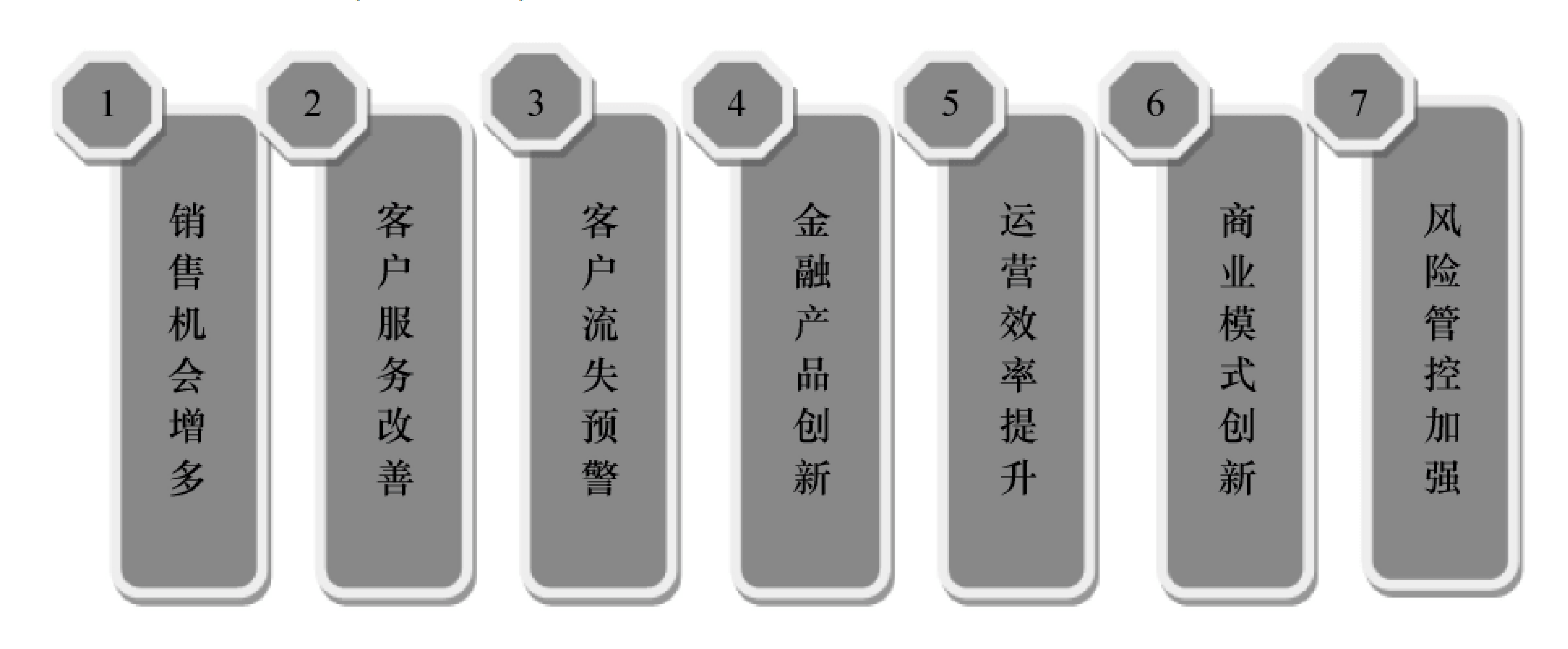


图 1.4 大数据在金融行业中的价值

1. 销售机会增多

金融企业掌握了海量的资金往来数据，再结合用户搜索行为、浏览行为、交易行为、评论历史、个人资料等数据，金融企业可以洞察消费者的整体需求，进而有针对性地进行产品生产、改进和营销。《纸牌屋》选择演员和剧情、百度基于用户喜好进行精准广告营销、阿里根据天猫用户特征包下生产线定制产品、亚马逊预测用户点击行为提前发货均是受益于互联网用户行为预测。

2. 客户服务改善

大数据的应用可以有效地改善客户服务。大数据不仅可以分析量化数据，还可以进行文本、语音分析。在客户体验方面，通过对交易数据、多渠道交互数据、社交媒体数据等的全面分析，帮助企业真正了解客户需求，并预测客户未来行为，从而为客户提供更好的

服务。在客户情感分析方面，通过对客服中心、社交媒体等数据的文本分析、语音分析，洞察客户情绪变化，分析客户的兴趣点、异常行为、意见、态度等，指导相关部门制定销售策略、市场策略等，并优化改进客户服务。

3. 客户流失预警

开发新客户往往比留住老客户要付出更高的成本。大数据技术的应用可以预警客户流失，减少客户流失率。利用大数据技术分析用户在整个相关产品里的使用行为的数据，识别可能流失的客户以及可能导致客户放弃的原因，如客户对产品不满意、对服务不满意、因为其他竞争对手等，以便企业及时采取策略，进行积极有效的改进。研究表明，客户在最终离开之前，很可能会持续关注或已经购买了竞争对手的产品，这些可以依据大数据进行探查。

4. 金融产品创新

大数据应用为金融行业突破传统金融产品带来了革新。高端数据分析系统和综合化数据分享平台能够有效地对接银行、保险、信托、基金等各类金融产品，使金融企业能够从其他领域借鉴并创造出新的金融产品。国内的数据挖掘最早基本也是基于授信所需要的分类挖掘算法而发展的。比如，金融贷款产品正在从抵押贷款向无抵押贷款演变，通过大数据应用建立信用评估机制，极大地提高了信用风险评级的及时性和准确性，抵押贷款模式正在逐步被信用贷款模式所取代。

5. 运营效率提升

在销售运营方面，金融机构能够通过现有客户的人际网络或业务网络，发现更多有价值的潜在客户，利用大数据的分析和预测模型，实现对客户消费模式和购买需求的分析，针对其个性需要展开精准营销，大大提升销售运营效率。在业务流程方面，通过大数据在存储和处理方面的优势，各种数据可被直接推送到需要这些信息的岗位，信息传递的中间环节被压缩，业务流程得到简化，从而带来巨大的效率提升空间。在资金需求预测方面，可以借助大数据构建资金需求预测模型，实现对资金需求的有效预算，帮助金融企业提高周转效率。

6. 商业模式创新

互联网金融和大数据技术正在对传统金融产生巨大冲击，大数据打破了信息不对称的局面，给金融商业模式带来了重大变化。一个很重要的表现形式是大数据的征信和网络贷款，可以根据企业行为数据计算出企业可能违约的概率，在这个基础上进行贷款，比如当前典型的阿里小贷。未来基于大数据的保险也是这样的，根据行为的数据进行保险差别的定价。比如，通过对人体的心率、体重、血脂、血糖、运动量、睡眠量等数据分析，预测客户的健康指数，帮助人身保险公司提高客户识别率，以此制定个性化的费率和承保方案。

7. 风险管控加强

由于金融的本质是对风险的控制和管理，这一特点决定了金融企业在风险管控方面的重视程度远远高于其他行业。风险管控是金融企业运营中的一个重要组成部分。风险发现得越早，挽回损失的概率越大。大数据的运用将大大有助于金融企业提升风险管控能力，通过对最底层交易数据的全面甄别和分析，使企业能够提高风险透明度，实现事前预警、事中控制。比如，大数据可以帮助银行建立动态的、可靠的信用系统，识别高风险客户以及各种交易风险，进而有效地进行防范和控制。

金融行业的业务范围是由客户、交易、资金、场所共同组成的联合体，任何一个要素的变化，都有可能带来意想不到的价值。

1.2 大数据应用领域

2012 年《纽约时报》的一篇文章标志着人类社会进入大数据时代，大数据影响着每一个人，并在可以预见的未来继续影响着整个人类和社会。大数据冲击着许多主要行业，大数据也在彻底地改变着我们的生活，未来大数据产业将会是一个很大的市场。目前，大数据已被广泛应用于各个行业，本书将主要为大家介绍大数据在商业、通信、医疗和金融这些应用比较早的领域中的应用。

1.2.1 商业

商业是大数据应用最广泛的领域。商业大数据的来源可分为两个方面：一方面是大交易数据，即商业交易产生的数据，包括商品数据、市场竞争数据、运营数据、销售数据、顾客关系数据和财务数据；另一方面是大交互数据，商业企业与顾客之间通过 POS、互联网、物联网、移动终端、智能终端、传感器和观测设备等产生的交互信息，主要包括社交网络数据、射频识别数据、时间和位置数据、文本数据和观测数据。大数据在商业中的应用可以归纳为以下 4 个方面(见图 1.5)。

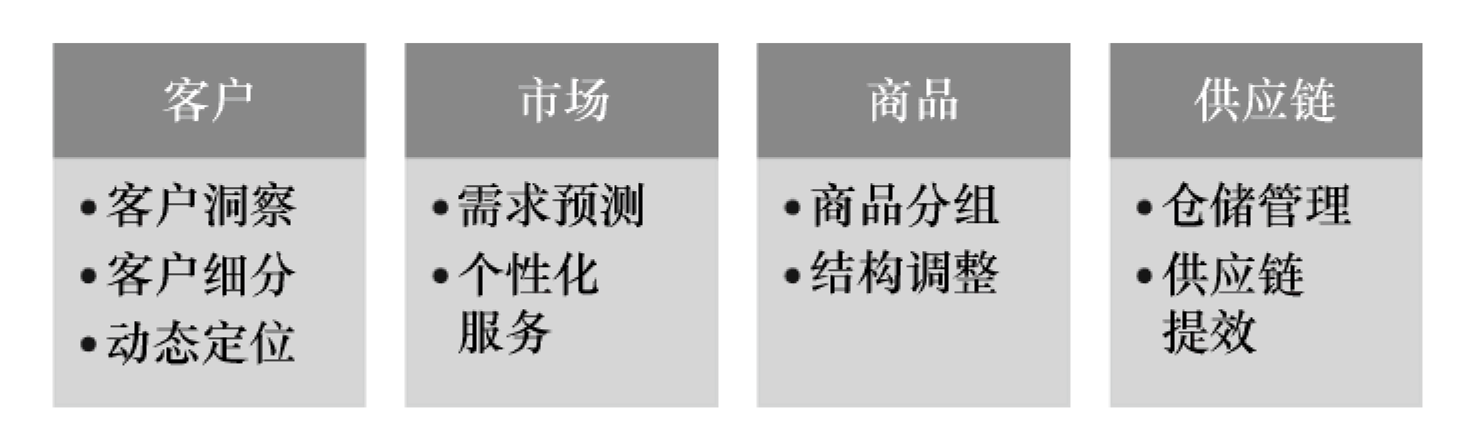


图 1.5　大数据在商业中的应用

1. 客户

在客户方面，大数据的应用主要包括客户洞察、客户细分和动态定位。①客户洞察。互联网、物联网等的顾客数据痕迹能真实而直接地反映消费者的性格、偏好和意愿。②客户细分。传统的以地理位置、人口统计特征为标准的划分被以爱好兴趣、生活方式、价值观、沟通方式为标准的数据化细分所替代；本质上讲，每个人的兴趣、爱好与需求都不

同，每个人都是一个细分市场，大数据正在使零售企业向“微市场”迈进，构建基于大数据的顾客购买行为模型，主动推荐个性化的产品和服务。③动态定位。零售业多来源、多格式数据的集成、分析与解释能力使数据的反馈与响应可在瞬间完成，快速识别消费者的购买决策和行为模式的变化趋势，及时准确地更新他们的偏好。

2. 市场

在市场方面，大数据的应用主要包括需求预测和个性化服务。①需求预测。通过对建构的大数据进行统计与分析，采取科学的预测方法，建立数学模型，使企业管理者掌握和了解零售行业潜在的市场需求，未来一段时间每个细分市场的产品销售量和产品价格走势等，从而使企业能够通过价格的杠杆来调节市场的供需平衡，并针对不同的细分市场来实行动态定价和差别定价。②个性化服务。根据客户的购买频次、兴趣点、忠诚度和流失的可能性预测客户的消费意愿，主动为其提供个性化的销售和关怀指导服务，调高销售额和利润率。

3. 商品

在商品方面，大数据的应用主要包括商品分组和商品结构调整。①商品分组。通过对代销记录信息的分析，可以发现购买某一种商品的顾客可能购买其他商品。这类信息可用于一定的购买推荐，或者保持一定的最佳商品分组布局，以帮助客户选择商品，刺激顾客的购买欲望从而达到增加销售额、节省顾客购买时间的目的。②商品结构调整。通过对销售数据和商品基础数据的分析，来指导企业商品结构的调整，加强所营商品的竞争能力和合理配置。

4. 供应链

在供应链方面，大数据的应用主要包括仓储管理和供应链提效。①仓储管理。通过对销售数据和库存数据的分析，决定各种商品的增减数量，确保正确的库存。②供应链提效。具体包括选择供应商，优化物流、现金流和配置人力资源等。利用大数据技术，优化整合供应链的各个环节，构建一个统一的供应链平台，各部门共享供应链平台的数据和服务，快速灵活地应对顾客消费变化，降低供应链成本，提高商品采购、仓储管理、物流配送和最终销售之间的运行效率。

大数据在零售商业中已有很多成功的应用案例。沃尔玛通过对消费者购物行为等这种非结构化数据的分析，了解顾客购物习惯，通过销售数据分析适合搭配在一起购买的商品，创造了啤酒与尿布的经典商业案例；淘宝数据魔方通过对消费者行为的分析帮助商家了解淘宝平台上的行业宏观情况、自己品牌的市场状况，据此进行生产、库存决策；美国折扣零售商 target 使用大数据分析，对顾客怀孕趋势进行评分，比较准确地预测了预产期，以此在每个孕期阶段为客户寄送相应的优惠券。在未来几十年，数据分析技术将不断地进步，商业领域将对组织、营销与管理进行突破性的创新。

1.2.2 通信

通信行业数据来源广泛，不仅涉及移动语音、固定电话、固网接入、无线上网等业

务，还会涉及公众客户、政企客户和家庭客户，同时也会收集到实体渠道、电子渠道、直销渠道等所有类型渠道的接触信息。通信行业发展至今积累了非常丰富的数据，既拥有财务收入、业务发展等结构化数据，还会涉及图片、文本、音频、视频等非结构化数据。目前，大数据在通信行业的应用还处于探索阶段，主要包括网络管理和优化、市场与精准营销、客户关系管理、企业运营管理和数据商业化 5 个方面，如图 1.6 所示。

网络管理和优化	市场与精准营销	客户关系管理	企业运营管理	数据商业化
•基础设施建设优化 •网络运营管理和优化	•客户画像 •关系链研究 •精准营销 •实时营销 •个性化推荐	•客服中心优化 •客户关怀与生命周期管理	•业务运营监控 •经营分析和市场监测	•营销洞察和精准广告 •大数据监测和决策

图 1.6 通信行业大数据应用

1. 网络管理和优化

网络管理和优化包括基础设施建设优化和网络运营管理和优化。①在基础设施建设层面，运营商运用大数据选择基站和热点的位置并有效地分配资源。例如，对话单和信令中用户的流量在时间周期和位置特点方面的分布进行分析，将 4G 基站和 WLAN 热点建立在 2G、3G 的高流量区域；与此同时，对已有基站的效率和成本建立评价模型，发现基站建设的资源浪费问题。②在网络运营方面，运营商可以利用大数据分析网络的流量和变化趋势及时调整资源配置，通过对网络日志进行分析优化网络，提升网络质量和利用率。③在网络优化方面，运营商可以运用大数据技术实时监控网络状况，对各个小区的网络数据进行综合分析，识别业务热点小区，依次设定网络优化的优先级，实现网络和用户的智能指配，提高投资效率。

2. 市场与精准营销

市场与精准营销包括客户画像、关系链研究、精准营销、实时营销和个性化推荐。①客户画像。运营商根据客户终端信息、地理位置、通话行为数据挖掘对客户群体进行分类，给每个客户打上行为和爱好标签，完善客户画像，有助于运营商深入了解客户的行为偏好和需求。②关系链研究。运营商可以运用客户资料和通话行为等数据分析客户交往圈，发现高流量用户，寻找营销机会，从而节约成本，提高营销效率。③精准营销。运营商可以通过大数据技术对用户终端的消费能力、消费偏好和近期特征事件进行分析，预测用户需求，精准匹配用户和通信相关业务，寻找合适的推送渠道、推送时间，实现精准营销。④个性化推荐。运营商可以通过对客户画像信息、终端信息、行为偏好等的分析，向客户提供定制化服务，优化产品设计和定价机制，实现个性化推荐和服务，提升客户体验。

3. 客户关系管理

客户关系管理包括客服中心优化、客户关怀和客户生命周期管理。①客服中心优化。

首先，运营商可以通过对客服中心积累的客户的呼叫行为和需求数据进行大数据分析，运用呼入客户行为数据和客户历史情况建立客服热线智能路径模型，预测客户的投诉风险，从而提升客服满意度。其次，根据语义分析，识别热点问题和客户情绪，通知相关部门进行优化。②客户关怀与客户生命周期管理。一是获取客户阶段，可以运用大数据技术挖掘和发现潜在客户。二是客户发展阶段，运用关联规则等数据挖掘方法进行交叉销售，促进客户消费。三是客户成熟阶段，利用大数据对客户群进行分类，实施精准营销，同时对不同客户进行个性化推荐。四是客户衰退阶段，采用预警模型预先发现高流失风险客户，做出相应的客户关怀。五是客户离开阶段，通过大数据挖掘高潜回流客户，推出客户感兴趣的业务，防止流失。

4. 企业运营管理

企业运营管理，包括业务运营监控、经营分析和市场监测。①业务运营监控。运营商可以运用大数据技术从网络、业务、用户等多个方面为运营商监控管道和客户运营情况。此外，还可以建立 KQI、KPI 等指标体系和异动智能监控体系，全面、及时、准确地监控业务运用情况。②经营分析和市场监测。运营商可以通过分析企业内部的业务和用户数据以及通过大数据技术采集的外部社交网络数据和市场数据，对业务和市场经营状况进行总结，主要包括经营日报、周报、月报、季报和年报。

5. 数据商业化

数据商业化是指企业通过自身拥有的大数据资产进行对外商业化，获得盈利。相比于国外，国内的数据商业化还处于探索阶段。数据商业化包括营销洞察、大数据监测和决策支撑服务。①营销洞察。美国电信运营商 Verizon 成立了专门的精准营销部门，主要用于提供精准营销洞察和商业数据分析服务。例如，在美国商家最为看中的营销场合，Verizon 对观众的来源进行了精确的数据分析，球队因此能够了解到观众对赞助商的喜好等。②大数据监测和决策。在客流和选址方面，西班牙电信成立了动态洞察部门开展大数据业务，主要为客户提供数据分析打包服务。该公司与市场研究机构 GFK 进行合作推出的产品“智慧足迹”通过完全匿名和聚合的移动网络数据，帮助零售商分析顾客来源和各商铺、展位的人流情况以及消费者特征和消费能力，并将洞察结果面向政企客户提供客流分析和零售店选址服务。在公共事业服务方面，法国电信运营商的通信解决方案部门承担了法国很多公共服务项目的 IT 系统建设，如法国高速公路数据监测项目，对其每天产生的记录进行分析就可以为行驶的车辆提供准确及时的路况信息，从而有效提高道路通畅率。

由于我国运营商的区域化运营，由各地区分公司分别存储通信企业的数据，而没有统一和整合，导致数据孤岛效应严重。因此，我国通信大数据仍然处于初级探索阶段。通信行业数据的整合和统一是大数据运用的重要一步。我国通信行业目前正着手准备这方面的工作，相信中国的通信行业大数据发展在互联网的竞争压力下会更快。

1.2.3 医疗

医疗行业拥有大量病例、病理报告、医疗方案、药物报告等。如果对这些数据进行整

理和分析，将会极大地帮助医生和病人。医疗行业大数据目前尚未统一收集起来，无法进行大规模应用。在未来，借助于大数据平台我们可以收集疾病的基本特征、病例和治疗方案以及病人的基本特征，建立针对疾病特点的数据库，帮助医生进行疾病诊断。医疗行业大数据来源如图 1.7 所示。

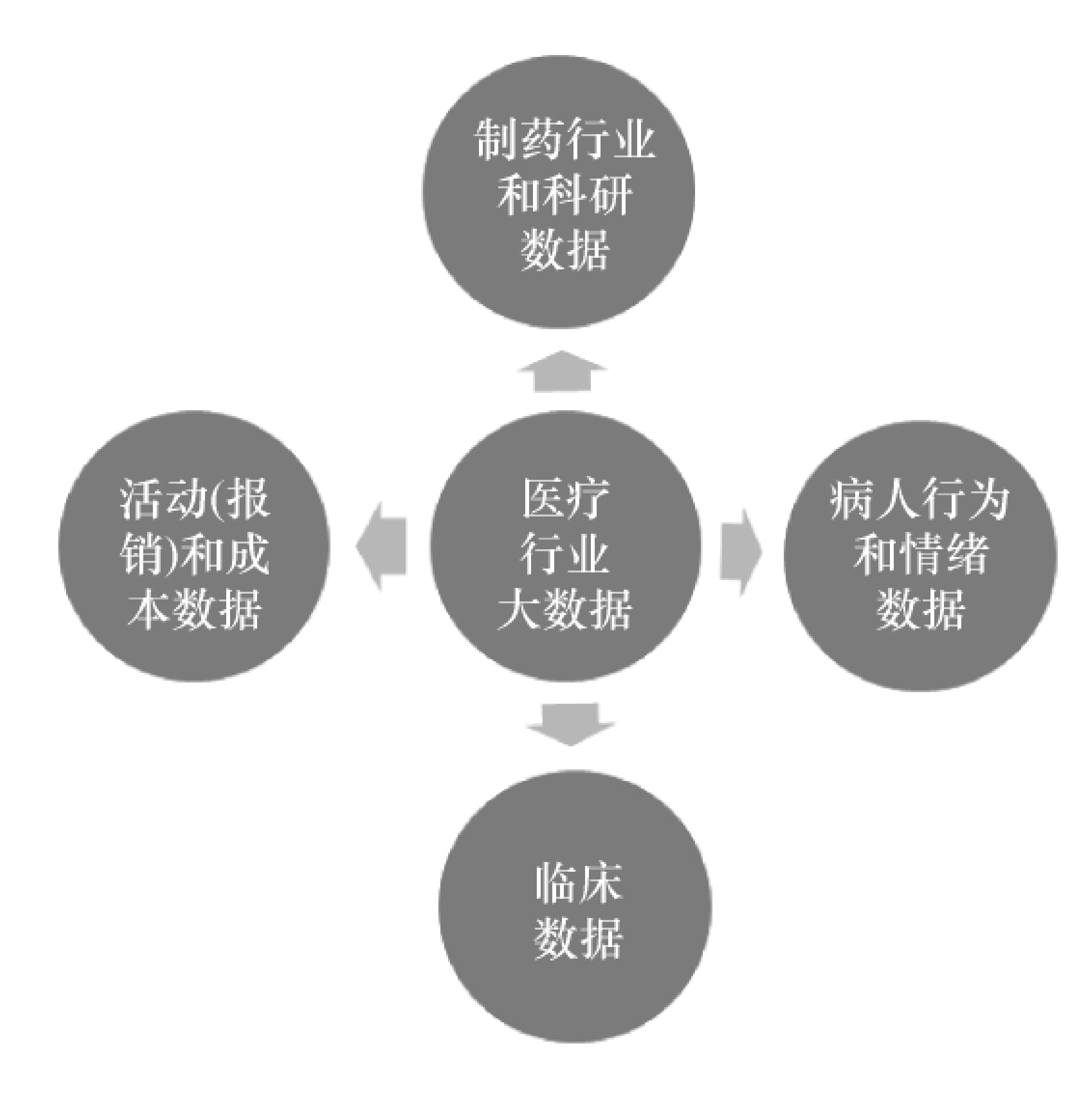


图 1.7　医疗行业大数据来源

大数据在医疗行业中的应用主要包括临床操作、付款/定价、研发、新的商业模式、公共健康这 5 个方面，如图 1.8 所示。

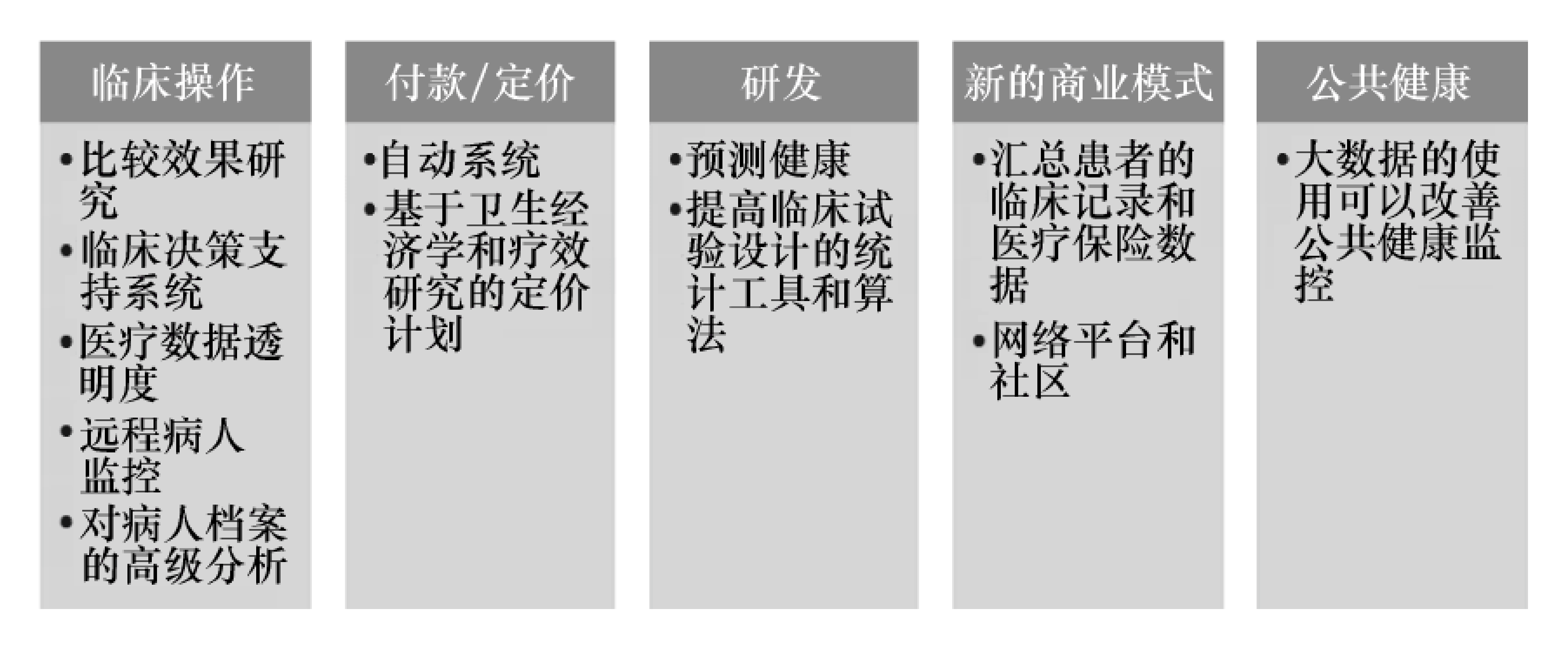


图 1.8　大数据在医疗行业中的应用

1. 临床操作

临床操作包括比较效果研究、临床决策交持系统、医疗数据透明度、远程病人监控和对病人档案的高级分析。例如，通过对病人的体征数据、费用数据和疗效数据在内的大型数据集进行精准分析，比较多种干预措施的有效性可以针对特定病人找到最有效和最具有成本效益的治疗方法；使用图像分析和识别技术，识别医疗影像(X 光、CT、MRT)数据，或者挖掘医疗文献数据建立医疗专家数据库，从而给医生提出诊疗建议；根据医疗服务提供方设置的操作和绩效数据集，可以进行数据分析并创建可视化的流程图和仪表盘，促进

信息透明，帮助病人做出更明智的健康护理决定，间接提高医疗服务的质量；从对慢性病人的远程监控系统收集数据，并将结果反馈给监控设备(查看病人是否遵从医嘱)，从而确定今后的用药和治疗方案；对病人档案的高级分析，确定各类疾病的易感人群，识别患病风险，使他们尽早接受预防性保健方案。

2. 付款/定价

付款/定价包括自动化系统、基于卫生经济学和疗效研究的定价计划。例如，利用自动化系统(机器学习技术)对索赔数据进行分析和挖掘，可以检测出索赔准确性，在支付发生前识别欺诈行为，避免重大的损失；利用数据分析横向医疗服务提供方的服务，并依据服务水平进行定价。

3. 研发

研发包括预测健康、调高临床实验设计的统计工具和算法、临床试验数据的分析、个性化治疗以及疾病模式的分析。例如，医药公司在新药物的研发阶段可以基于药物临床试验阶段之前的数据集及早期临床阶段的数据集，及时地预测临床结果；在临床试验阶段通过统计工具和算法挖掘病人数据，评估招募患者是否符合试验条件，加快临床试验进程；根据临床试验数据和病人记录确定药品更多的适应证以及从中发现副作用；通过对大型数据集(如基因组数据)的分析发展个性化治疗；对疾病的模式和趋势分析，帮助医疗产品企业制定战略性的研发投资决策，优化研发重点和配备资源。

4. 新的商业模式

新的商业模式包括汇总患者的临床记录和医疗保险数据集、网络平台和社区。例如，汇总患者的临床记录和医疗保险数据集，并进行高级分析，将提高医生和医药企业的决策能力。在医生诊断病人时可以参考病人的疾病特征、化验报告和检测报告，参考疾病数据库来快速帮助病人确诊，明确定位疾病。在制定治疗方案时，医生可以依据病人的基因特点，调取相似基因、年龄、人种、身体情况相同的有效治疗方案，制定出适合病人的治疗方案，帮助更多人及时进行治疗。同时这些数据也有利于医药行业开发出更加有效的药物和医疗器械。另一个潜在的大数据启动的商业模型是网络平台和大数据，这些平台已经产生了大量有价值的数据：包括病人的问诊数据、医生的学习习惯等。

5. 公共健康

大数据的使用可以改善公众健康监控。公共卫生部门可以通过覆盖全国的患者电子病历数据库，快速检测传染病，进行全面的疫情监测，并通过集成疾病监测和响应程序，快速进行响应。这将带来很多好处，包括医疗索赔支出减少、传染病感染率降低，卫生部门可以更快地检测出新的传染病和疫情。通过提供准确和及时的公众健康咨询，将会大幅提高公众健康风险意识，同时也将降低传染病感染风险。所有的这些都将帮助人们创造更好的生活。

大数据将会对医疗行业产生巨大的影响和推动，它可以揭露健康的影响因素，将最合适的治疗方式推荐给患者；能够促进新的发现，优化治疗结果和削减开支。但目前大数据

医疗也面临着患者隐私安全、海量数据收集难题、区域医疗共享以及技术方面的挑战。随着信息化技术的发展，这些问题将逐步得到解决。可以预见，在不久的未来，大数据的应用将渗透到医疗应用的更多领域。

1.2.4 金融

在国外，大数据在金融行业中的应用开展较早。例如，美国银行运用客户点击数据集为客户提供特色服务，包括有竞争性的信用额度；花旗银行运用 IBM 沃森电脑为财富管理客户推荐产品。中国金融行业大数据应用主要在近几年运用较为广泛，很多金融机构建立了大数据平台，采集和处理金融行业的交易数据，主要应用于金融行业的营销、服务、运营和风控 4 个方面，如图 1.9 所示。

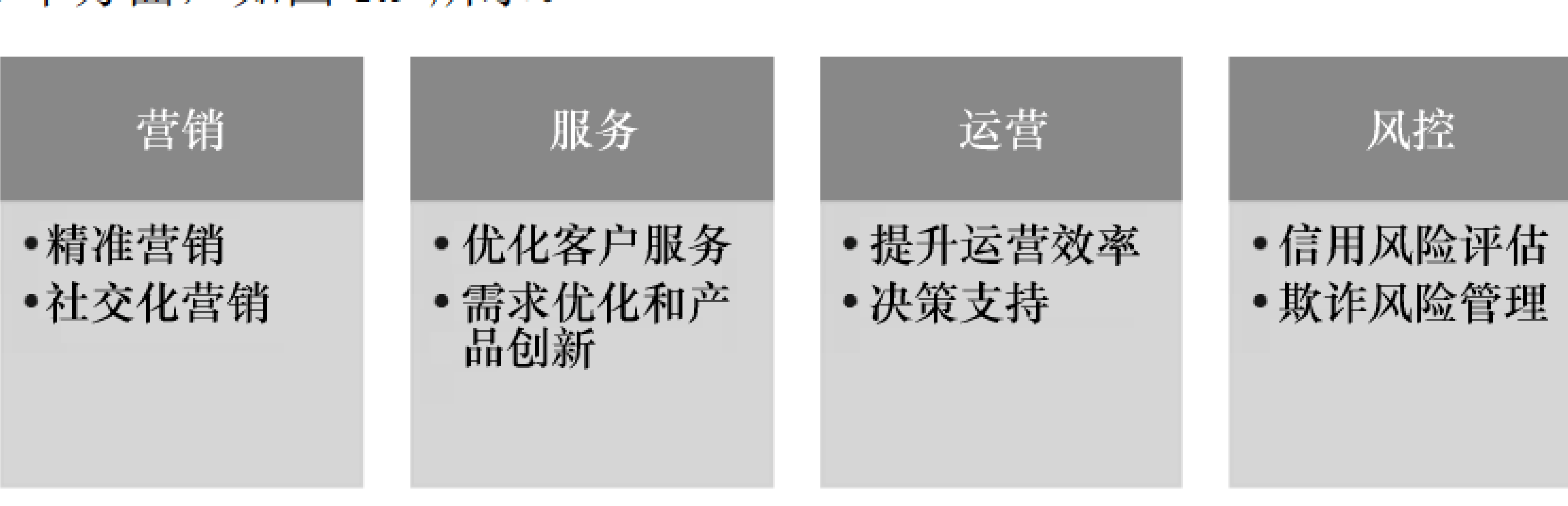


图 1.9　大数据在金融行业的应用

1. 营销

1)　精准营销

精准营销是指根据客户的消费偏好和消费能力确定目标客户，推荐个性化产品。例如，银行对客户刷卡、存款取款、银行转账、微信评论等行为数据进行整理和分析，定期向客户推送广告信息，包括客户可能感兴趣的产品和优惠信息；信用卡中心可以利用大数据追踪热点消息，针对特定人群提供产品，如热映电影、娱乐活动、美食饮品等；证券公司可以通过大数据分析为特定企业提供融资融券产品；保险公司可以根据大数据定制有针对性的保险产品。精准营销的具体流程如图 1.10 所示。

基础平台
•根据用户标签需求建立各个标签数据来源的数据基础平台

数据建模
•通过数据挖掘、多维分析、统计技术建立各种模型进行标签计算输出

数据更新
•数据来源更新，周期性地进行模型学习更新，同步更新用户画像数据

精准营销
•用户画像结果数据应用到精准营销，定向推荐金融产品

图 1.10　精准营销流程

2)　社交化营销

社交化营销是指利用社交平台的数据资源，结合大数据分析进行营销。金融行业可以开展成本较低的社交化营销，凭借开放的互联网平台，通过对大量的客户需求数据进行分

析，进行产品和渠道推广。然后依据互联网社交平台反馈的用户数据，评价营销方案的可行性，利用口碑营销和病毒式传播来帮助金融行业快速进行产品宣传、品牌宣传、渠道宣传等。

2. 服务

1) 优化客户服务

银行可以根据大数据分析，在节假日问候客户，为客户提供定制服务，预知网点客户的未来资金需求，提前进行预约，提高客户体验；私人银行还可以通过大数据分析，代理客户参与金融市场投资，获取超额利润，优化客户服务。证券公司可以通过大数据分析，快速推出相应的行业报告和市场趋势报告，以利于投资者及时了解热点，优化客户服务；保险公司可以根据大数据预测为客户提前提供有效服务，改善客户体验，同时增加商业机会。

2) 需求分析和产品创新

银行可以从职业、年龄、收入、居住地、习惯爱好、资产、信用等各个方面对客户进行分类，依据其他的数据输入维度来确定客户的需求并定制产品。银行还可以依据企业的交易数据来预测行业发展特点，为企业客户提供金融产品服务。保险行业可以依据外部数据导入，根据热点词汇来判断市场对保险产品的需要。证券公司也可以依据外部数据判读投资者喜好，来定制投资产品，进行产品创新。

3. 运营

1) 提升运营效率

大数据可以展现不同产品线的实际收入和成本，帮助银行进行产品管理。同时，大数据为管理层提供全方面的报表，揭示内部运营管理效率，有利于内部效率提升。大数据可以帮助市场部门有效监测营销方案和市场推广情况，提高营销精度，降低营销费用。大数据可以通过展现风险视图来控制信用风险，同时加快信用审批。大数据可以帮助保险行业快速为客户提供保险方案，提高效率，降低成本。证券行业也可以利用大数据动态提供行业报告，快速帮助投资人。

2) 决策支持

大数据可以帮助金融企业，为即将实施的决策提供数据支撑，同时也可以依据大数据分析归纳出规律，进一步演绎出新的决策。基于大数据和人工智能技术的决策树模型将会有效帮助金融行业分析信用风险，为业务决策提供有力支持。金融行业新产品或新服务推向市场前，可以在局部地区进行试验，大数据技术可以对采集的数据进行分析，通过统计分析报告为新产品的市场推广提供决策支持。

4. 风控(风险控制)

1) 信用风险评估

银行可以利用大数据增加信用风险输入维度，提高信用风险管理水平，动态管理企业和个人客户的信用风险。建立基于大数据的信用风险评估模型和方法，将会提高银行对中小企业和个人的资金支持。个人信用评分标准的建立，将会帮助银行在即将到来的信用消

费时代取得领先地位。基于大数据的动态的信用风险管理机制，将会帮助银行提前预测高风险信用违约时间，及时介入，降低违约概率，同时预防信用欺诈。

2) 欺诈风险管理

信用卡公司可以利用大数据及时预测和发现恶意欺诈事件，及时采取措施，降低信用欺诈风险。保险公司可以利用大数据发现恶意投保和索赔事件，降低欺诈带来的经济损失。银行可以基于大数据建立防欺诈监控系统，动态管理网上银行、POS、ATM 等渠道的欺诈事件。大数据提供了多维度的监控指标和联动方式，可以弥补和完善目前反欺诈监控方式的不足。特别在识别客户行为趋势方面，大数据具有较大的优势。

金融行业的数据丰富，通过对客户信息、交易信息、资产信息、信用信息等数据的采集和整理，结合外部数据分析，可以有效帮助金融行业进行精准营销、提高运营效率、优化客户服务、进行产品创新、提高信用风险和欺诈风险管理水平，为决策提供有效支持。但在大数据时代，金融行业也面临着诸如自身技术、信息安全、金融监管等方面的挑战，相信随着大数据技术的发展，这些问题会逐步得到解决。

1.3 大数据金融的内涵、特点与优势

1.3.1 大数据金融的内涵

大数据金融是指运用大数据技术和大数据平台开展金融活动和金融服务，对金融行业积累的大数据以及外部数据进行云计算等信息化处理，结合传统金融，开展资金融通、创新金融服务。具体来说，大数据金融通过收集和整合海量的非结构化数据，运用大数据、互联网、云计算等信息化方式，对客户消费数据进行实时分析，可以为金融企业提供客户全方位信息，通过分析和挖掘客户的交易和消费信息掌握客户的消费习惯，准确预测客户行为，提高金融服务平台新的效率以及降低信贷风险。

金融行业的大数据大致分为以下 3 类。

(1) 传统的结构化数据，如各种数据库和文件信息等。

(2) 社交媒体为代表的过程数据，涵盖了用户偏好、习惯、特点、发表的评论，朋友圈之间的关系等。

(3) 日益增长的机器设备以及传感器所产生的数据，如柜面监控视频、呼叫中心语音、手机、ATM 等记录的位置信息等。

根据金融行业的分类，可以将大数据金融细分为大数据银行、大数据保险和大数据证券。差异化车险定价是典型的大数据保险形式之一，是指保险行业利用驾驶信息来确定车险价格，良好驾驶习惯的车主，其车险价格就较低，反之车险价格就较高；信用卡自动授信是典型的大数据银行的应用，银行根据用卡客户数据确定是否授信以及计算信用额度；机器人投资是大数据证券的创新模式之一，证券公司根据股价的影响因素建立模型，自动选择股票或寻找交易时机，在适当的风控模型下建立机器人投资云交易模式。

1.3.2 大数据金融的特点

大数据金融与传统金融相比，存在如下几个方面的特点。

1. 呈现方式网络化

在大数据金融时代，大量的金融产品和服务通过网络呈现，如支付结算、网络借贷、P2P、众筹融资、资产管理、现金管理、产品销售、金融咨询等都将主要通过网络实现。网络也包括固定网络和移动网络，其中移动网络将逐步成为大数据金融服务的主要途径。

2. 风险管理有所调整

在风险管理理念上，财务分析(第一还款来源)、可抵押财产或其他保证(第二还款来源)重要性将有所降低。交易行为的真实性、信用的可信度通过数据的呈现方式将会更加重要，风险定价方式将会出现革命性变化。对客户的评价将是全方位、立体的、活生生的，而不再是一个抽象的、模糊的客户构图。基于数据挖掘的客户识别和分类将成为风险管理的主要手段，动态、实时的监测而非事后的回顾式评价将成为风险管理的常态性内容。

3. 信息不对称性降低

在大数据金融时代，金融产品和服务的消费者和提供者之间的信息不对称程度会大大降低。对某项金融产品(服务)的支持和评价，消费者也可实时获知。

4. 金融业务效率提高

大数据金融的许多流程和动作都是在线上发起和完成的，有些动作是自动实现的。在合适的时间、合适的地点，把合适的产品以合适的方式提供给合适的消费者。同时，强大的数据分析能力可以将金融业务做到极高的效率，交易成本也会大幅降低。

5. 金融企业服务边界扩大

首先，对于单个金融企业，最适合扩大经营规模，由于效率提升，其经营成本必然随之下降。金融企业的成本曲线形态也会发生变化，长期平均成本曲线的底部会更快来临，也会更平坦、更宽。其次，基于大数据技术，金融从业人员个体服务对象会更多，即单个金融企业从业人员会有减少的趋势，或至少其市场人员有降低的趋势。

6. 产品是可控的、可接受的

通过网络化呈现的金融产品，对消费者而言，是可控、可接受的。产品可控是指在消费者看来，其风险是可控的。产品可接受是指在消费者看来，首先其收益或成本是可以接受的；其次，产品的流动性是可以接受的；最后，基于金融市场的数据信息，消费者认为其产品也是可以接受的。

7. 普惠金融

大数据金融的高效率性及扩展的服务边界，使金融服务的对象和范围也大大扩展，金融服务也更接地气。例如，极小金额的理财服务、存款服务、支付结算服务等普通老百姓

都可以享受到，甚至极小金额的融资服务也会普遍发展起来，金融深化在大数据金融时代可以完全实现。

1.3.3 大数据金融相对于传统金融的优势

传统金融对数据的重视程度不高，数据分析技术落后，大数据技术的应用相对缺乏。相比传统金融，大数据金融具有如下优势。

1. 放贷快捷，精准营销个性化服务

大数据金融建立在长期的大量的信用及资金流的大数据基础之上，在任何时点都可以通过计算得出信用评分，并采用网上支付方式，实时根据贷款需要及其信用评分等数据进行放贷。大数据金融根据企业不同的生产流程和信用评分进行放贷，不受时空限制，较好地匹配了企业的期限管理，解决了企业的流动性问题。此外，大数据金融还可以针对每一家企业的个性化融资需求做出不同的金融服务且快速、准确、高效。

2. 客户群体大，运营成本低

传统金融主要是以人工为主体参与审批，大数据金融是以大数据云计算为基础，以大数据自动计算为主，不需要大量人工，成本较低，不仅可以针对小微企业提供金融服务，还可以根据企业生产周期灵活调整贷款期限。大数据金融整合了碎片化的需求和供给，将服务领域拓展至更多的中小企业和中小客户，更大程度地降低了大数据金融的运营成本和交易成本。

3. 科学决策，有效风控

网络借贷平台或供应链聚集了信息流、物流和资金流，其借贷信息都累积在大数据金融库持久闭环的产业上下游内部，贷款方对产业运作和风险点比较熟悉且容易掌控，有利于风险的防范和预警。大数据金融可以根据这些交易借贷行为的违约率等相关指标估计信用评分，运用分布式计算做出风险评估模型，解决信用分配、风险评估、授权实施以及欺诈识别等问题。通过以大数据金融为基础的风控科学决策，有效地降低了不良贷款率。

大数据金融相比于传统金融有无可比拟的优势。企业可以通过大数据金融对商业模式和盈利模式加以创新，获得在产业链中的核心地位。大数据金融带来的技术革新和金融创新不仅能支持中小企业的发展，还能促进我国经济结构调整和转型升级。因此，大数据金融战略是企业和国家的战略选择。

1.4 大数据带来金融业大变革

随着计算机技术和互联网的发展，金融行业的数据采集能力逐步提高，存储了大量时间连续、动态变化的金融数据。相比于其他行业，大数据对金融业更具有潜在价值。麦肯锡的研究表明，金融业在大数据价值潜力指数中排名第一。伴随着大数据的应用、技术革新以及商业模式的创新，金融交易形式日趋电子化和数字化，具体表现为支付电子化、渠

道网络化、信用数字化，运营效率得到极大提升。银行、保险、证券等传统金融行业迎来了巨大的变革。

1.4.1 大数据带来银行业大变革

近几年，大数据高速发展，使得银行业的客户数据、交易数据、管理数据等均呈现爆炸式增长。据中国银联公开数据显示，全国仅“银联”银行卡的发行量目前就接近 40 亿张，每天有近 600 亿元的交易通过银联的银行卡进行。如果再加上开户信息数据、银行网点和在线交易的各种数据，以及金融系统自身运营的数据，目前国内银行每年上升的数据能达到数十 PB。数据海量增长为银行业带来了机遇和挑战，其服务与管理模式已逐步发生改变。

1. 电子商务平台和电子银行

2012 年开始，多家商业银行开设了自己的电子商务平台，其中以建设银行、中国银行、交通银行的规模最大。这些购物网站与其他电商并没有太大的差别，包括吃穿住行等方面。另外，还有一些商业银行使用其他途径参与电商。商业银行挑战电商市场，其目的并不在于网上商城的营业收入，而在于扩展客户数据，使客户数据立体化，以了解客户消费习惯、消费能力、兴趣数据、风险偏好等进行客户画像的构建，预测客户行为，进行差异化服务。

银行大力投资改革网上银行业务。相比阿里巴巴、腾讯等跨界者，银行在资金、风险管理能力、人才储备等方面具备优势。国内多家银行大力投资于网上平台、推出网上服务，进行多元化创新，为发展自有互联网金融业务奠定基础。目前，商业银行的网上服务包括传统银行业务、电子商务与移动支付，以及 P2P 等新兴业务等。

2. 客户个性营销

随着利率市场化和民营银行设立预期的加剧以及互联网金融的兴起，银行业竞争日益激烈，利差进一步缩窄，银行纷纷进行发展模式的战略转型。实现战略转型目标要求银行必须可靠、实时掌握客户的真实需求，全面完整描述客户的真实面貌。大数据的发展为上述需求提供了技术条件，通过广泛收集各渠道、各类型的数据，使用大数据技术整合各类信息、还原客户真实面貌，可以帮助银行切实掌握客户的真实需求，并根据客户需求做出快速应对，实现精准营销和个性化服务。例如，新加坡花旗银行根据客户的刷卡时间和地点，结合客户的购物、餐饮习惯等个人虚拟性，可以精确地向客户推荐商场及餐厅优惠信息。

3. 银行风险管理

风险管理是银行的生命线。以往银行在进行信用风险管理时，主要依据客户的会计信息、客户经理的调查、客户的信用记录以及客户抵押担保情况等，通过专家判断进行决策。大数据技术的应用使银行的风险管理能力大幅提高。一方面，通过多种传感器、多个渠道采集数据，使银行更全面、更真实、更准确、更实时地掌握借款人的信息，有效降低信息不对称带来的风险。另一方面，利用大数据技术可以找到不同变量之间的关联关系，

形成新的决策模型，使决策更加准确、统一和合理。银行利用大数据能够创新风险决策模式，赢得新客户，形成利润增长点。如图 1.11 所示是大数据风险管理的基本步骤。

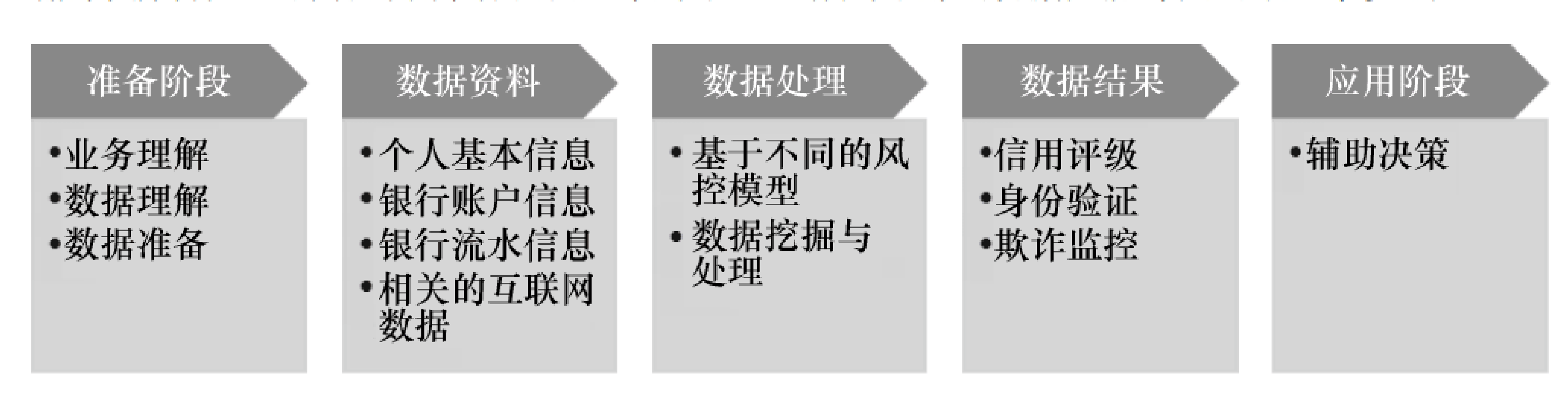


图 1.11　大数据风险管理的基本步骤

1.4.2　大数据带来保险业大变革

大数据与保险业具有天然的关联性。保险经营的核心基础是大数法则，如保险生命表就是以十万人为组来进行测算。无论是财产保险的概率事件，还是寿险的概率生命期，都是由大量数据分析获得的规律。长期以来，保险业通过上门、柜面、信函、电话、短信、微信等多种方式，已经积累了大量的客户交互数据。近年来兴起的互联网保险也成为保险业收集数据的新平台。据统计，国内大型保险公司每年新增的数据量达到 PB 级。在全球保险大数据应用市场中，主要领域包括客户行为分析、承保定价、互联网数据分析、市场渠道分析、风险建模、预测分析、商业决策、欺诈侦测等。

1. 承保定价

在大数法则下，保险产品的定价主要是基于样本数据的分析。大数据时代，保险定价是基于社会和全体数据，不仅包括保险公司存储的客户数据，还包括整个互联网上的数据，如来自社交网络上的文字、图片或者视频信息。这将颠覆传统保险精算的理论和技术，推动保险商业模式的革命性和突破性创新。车险将采用差别定价模式，生命表也将发生更新换代式的变革，所有的投保人将获得一个公平的保险价格。例如，保险公司可以通过数据分析，掌握客户车辆主要用途、基本行车路线、路途的风险程度、驾驶习惯等风险状况，以此评估客户车辆的风险指数，进而制定差别费率，对于风险低的客户降低费率，对于风险高的客户提高费率甚至拒绝承保。

2. 精准营销

传统的广告宣传手段是采用传统媒体，如电视、广告牌等，每个用户看到的广告一样，若该用户没有相关需求，广告也就没有效果。大数据时代的保险营销不是针对所有群体的一个广告及营销手段，而是实施精准营销。精准营销是通过分析客户行为，制定相应的销售与服务策略，把合适的产品或服务，以合适的价格，在合适的时间，通过合适的渠道，提供给合适的客户。大数据技术的应用，可以帮助保险公司完成寻找目标客户、挖掘客户潜在保险需求等任务。大数据营销使保险公司的客户营销策略更为精确直接，避免以往常见的逐户、陌生拜访、陪同拜访现象，也避免了和同业竞争对手直接碰撞。相比开拓新客户，大数据营销对原有客户购买力的深度挖掘和忠诚度培养具有重要意义。

国内友邦保险开通了网上服务自助平台及微信服务平台，开发客户地图等系统工具，帮助销售人员科学管理和分析客户在不同人生阶段的保障、理财需求。试验的 O2O 模式，已经初见成效。线上的精准定位和前期需求的挖掘，与线下高效的销售流程相结合，有效提升了客户转化率，也为企业创造了价值。

3. 欺诈识别

保险欺诈，尤其是健康保险领域的欺诈，具有专业性、隐蔽性等特点。保险公司主要是依靠一些固定标准和列配人员的经验，来判断是否存在保险欺诈。由于缺乏行业内协作机制和共享的信息平台，调查的质量主要是依赖于理赔人员的个人素质以及公安机关的合作情况。从本质上看，欺诈是由双方信息不对称所导致的，大数据能够弱化部分不对称的信息，建立高效的反欺诈鉴别机制。

为了防范健康保险交易中诈骗的发生，美国各州在建立全民医疗保险的网络销售平台时，附加建立专业软件平台，用于自动识别和侦破可疑的健康保险索赔数据。在国内，全国各地保险公司正在积极建设客户理赔信息即时共享机制、完善统一的欺诈风险信息库，以及广泛的异地协查网络，积极实现商业保险与社会保险之间的实时对接，扩大共享范围，提高支撑识别保险欺诈的数据质量。

1.4.3 大数据带来证券业大变革

随着 A 股市场全面放开一人一户限制、证券经营牌照将会向互联网公司放开，面对居民财富迅速增长和其对理财产品多样化的需求，证券公司受到来自行业内外部的双重压力，当前它们正在进行业务转型。传统 IT 基础设施环境已经无法满足证券公司对转型和创新战略的要求。随着大数据时代的到来，对于证券公司，数据驱动的创新平台的建设为即将到来的业务差异化竞争提供了强有力的技术支持。相比于银行业和保险业，证券行业的大数据应用相对较晚，正处于起步阶段，目前大数据主要应用于个性化服务、量化投资和股价预测。

1. 个性化服务

券商作为金融中介的职能在信息技术的冲击下将有所改变。在大数据背景下，券商将有能力快速收集、传导大量的高质量信息，以设计出符合客户需求的产品组合，并不断根据客户偏好的改变而调整。同时，通道中介服务深陷同质竞争，争夺焦点必然落到价格上。但是如果标准化同质服务不再能够给券商带来正常利润，最优选择要么是从竞争中彻底退出，要么是转变经营思路，将通道业务转变成包含增值服务的金融服务。

大数据在加强风险管控、精细化管理、服务创新等转型中别具现实意义，是实现向信息化券商转型的重要推动力。首先，大数据能够加强风险的可审性和管理力度。其次，大数据能够支持精细化管理。当前，中国证券业以客户为中心的管理改革已经起步，必然会对券商提出精细化管理的新要求。再次，大数据支持服务创新，能够更好地实现“以客户为中心”理念，通过对客户消费行为模式进行分析，提高客户转化率，开发出不同的产品以满足不同客户的市场需求，实现差异化竞争。

在过去的 10 年里，越来越多的证券公司采用数据驱动的方法进行有针对性的服务来降低风险和提高业绩。通过执行特殊的数据分析程序来对一系列资料进行收集、存储、管理和分析大数据集，识别关键业务，以便给客户提供更好的决策。可利用的金融数据源包括股票价格、外汇和衍生品交易、交易记录、高频交易、无结构化新闻和文本以及隐含在社会媒体和网络中的消费者信心和商业情绪。

2. 量化投资

随着互联网的发展，证券行业已经进入一个大数据信息海洋的云时代。“光大证券乌龙事件”也彻底表明了在一般投资者面对操盘的是冰冷的电脑方程式，证券的数据模型更加复杂多样，数据的总量和种类都有着非常大的突破。“光大黑天鹅事件”或许只是 A 股市场此类事件的开始，因而针对此类事件的预警变得格外重要。大数据技术是预防“黑天鹅”的重要手段。大数据在处理证券数据时能加深对数据本身，主力资金和散户资金，以及散户和主力的行为、轨迹，主力和散户之间、主力和市场之间的关系、散户和市场之间的关系等多重关系的理解。如果能把这些数据使用好，包括数据、数据挖掘能力、算法、平台等，就能够很好地增加投资胜率。量化投资策略在欧美等发达国家的金融市场发展已经相对成熟，行业竞争越来越激烈。

量化投资由于其巨大收益，是大数据最早应用的领域，而其也符合大数据最重要的三大思维变革。随着互联网和移动互联网带来的信息化革命，个人投资者将能够轻松使用大数据获得实证支持，降低交易策略风险，投资能力将大幅提升。大数据让科技公司第一次有机会能够挑战传统的金融分析师和交易员，利用对各种全体数据的量化、重组和整合，低成本地建立针对各个市场、面向不同用户的交易策略，让投资者能够科学稳定地在全球市场投资。因此，大数据时代对金融投资的革命不仅仅是未来的趋势，而是正在实现的现实，谁能做到这一点，谁就能引领证券投资的未来。

3. 股价预测

传统的股票价格预测是利用股票形态分析理论对股票未来走势的方向和可能性做出预测，这种方法是从海量的历史数据中寻找和某只股票当前趋势相同或相似的趋势，并根据历史趋势判断未来股票价格。股市是个复杂的系统，仅仅根据历史数据进行预测比较片面，不一定准确。在大数据时代，通过网络产生的搜索数据、互动数据等也可以用来预测股市活跃度和股价走势变化。互动数据反映了投资者对某只特定股票的喜好与厌恶，可以简单描述为对股票的操作是持有还是卖出；搜索数据则代表投资者对某只股票的兴趣和关注点，关注度高意味着消息的影响力大。市场本身带有主观判断因素，投资者的情绪会影响投资行为，而投资行为直接影响资产价格。例如，英国对冲基金(Derwent Capital Markets)是基于社交网络建立的对冲基金，该基金通过分析 Twitter 上的数据内容感知市场情绪，依据对市场情绪数据的分析进行股价预测，进而指导投资者投资。此外，IBM 使用大数据信息技术成功开发了经济指标预测系统。借助该预测系统，可通过统计分析新闻中出现的单词等信息来预测股价等走势。这种经济指标预测系统首先从互联网上的新闻中搜索与“新订单”等与经济指标有关的单词，然后结合其他相关经济数据的历史数据分析与股价的关系，从而得出预测结果。

1.4.4 大数据带来征信行业大变革

传统征信包括线下的金融征信体系、社会征信体系、商业征信体系以及线上某一层级数据的单一分析的 IT 征信。而大数据征信数据来源更广泛，不仅包括上述征信体系，还包括利用互联网手段工具挖掘的电子商务、社交、网络行为等特征信息。随着社会经济的飞速发展，征信业所收集、存储、处理的信息数据量呈现爆炸式增长，其必然也会进入大数据时代。在大数据时代，大数据思想和技术以其自身的优势必将为征信业提供新的发展机遇，为征信数据、征信服务、数据采集、征信产品等带来一系列变革。

1. 征信数据

大数据时代的到来使得征信数据来源更为广泛，征信数据类型更为多样。在数据来源上，传统的征信数据主要来源于个人或者机构的借贷、赊购、担保、租赁、保险、信用卡等活动，这些活动中产生的行政处罚信息、缴纳各类社保和公共事业费用信息等都是征信数据。在大数据时代，征信数据更多的是来源于线上，互联网公司(如淘宝、京东等)通过客户网上的交易记录、评价等信息还有社交网络信息更加真实完整地了解客户的信用状况。在数据类型上，大数据技术使得征信数据不再限于数字、字符这些结构化数据，还包括图片、音频、视频等非结构化数据。例如，交通银行信用卡中心通过智能语音分析技术，提炼出隐藏在音频数据中的客户信息进行分析应用，每天的数据处理量达到 20GB。

2. 征信服务

在大数据时代，征信机构的服务更加及时、高效、全面。例如，在营销服务方面，征信机构运用大数据技术对客户相关数据信息进行收集，勾勒客户画像，从多个方面对客户群体进行细分，从而提供差异化服务，使得营销服务更具有针对性和有效性。在客户维护方面，大数据技术可以帮助征信机构更加便捷、及时、有效地收集和分析客户对征信产品和服务效果的需求，及时反馈客户提出的问题和建议，从而提升客户忠诚度。与此同时，还可以运用大数据技术对客户使用服务的相关数据和征信机构所流失客户的相关数据进行挖掘分析，有助于预测发现可能流失的客户，从而及时对客户维护策略加以改进，保证客户群体的稳定。

3. 数据采集

征信机构传统的数据采集手段因机构性质不同而有差异。一种是公共征信机构，一般是由中央银行经营管理，金融机构(如商业银行、信用卡公司等)被强制要求定期向中央银行报送借款人的相关数据和信息。另一种是私人征信机构，独立于政府和大型金融机构之外，通常通过协议或者合同的方式规范数据采集，其数据的主要来源有提供信息服务的金融机构信贷信息、政府平台公布的公共记录等。而在大数据时代，通常是采用人们生活中含有内建芯片、传感器、RFID(无线射频芯片)等具有电子神经的感知设备产品收集数据信息。这些设备与计算机连接以后，可以随时随地对人们生活产生的各种数据进行收集，所收集的数据内容更加丰富，数据类型更加多样。

4. 征信产品

传统的征信产品主要包括信用报告、信用评分、信用评级、信用风险管理类产品。在大数据时代，大数据技术有助于提升征信产品的质量，推动征信产品的创新，扩展产品服务范围，促进征信业的发展。例如，在征信产品推销方面，可以运用大数据技术对客户的生活习惯等数据进行挖掘分析，预测客户的潜在需求，有针对性地为客户推销相应的征信产品。在征信产品的改进方面，大数据时代的信用报告可以结合客户的生活习惯、性格特点、财务状况、兴趣爱好等信息数据综合评判个人信用状况。与此同时，征信产品的形式也将更加多样化，不仅可以是上报的报表、可视化的图表、详细的可视化分析，还可以是简单的微博或视频信息等。此外，大数据技术的应用能够使得信用评分和信用评级更加准确合理。

1.4.5 互联网金融中的大数据应用

近几年，互联网金融迅速发展，并不断出现新的模式和应用，但其本质还是属于金融范畴。互联网金融自然产生大数据，它是大数据应用最为广泛的领域。其核心是数据，互联网金融业竞争力的强弱未来将取决于数据的规模、数据的有效性、数据的真实性、数据分析和应用的能力。其中，大数据技术是互联网金融的重要技术支撑。人们在网上活动的信息都会形成数据，运用大数据技术对数据进行收集、整理、挖掘、分析和深度应用，从而实现互联网金融产品、技术、营销和风险的创新管理。目前，互联网金融的大数据应用包括精准营销、风险管理、信用评价等。互联网金融方兴未艾，相信还会不断出现新的应用。

1. 精准营销

大数据的应用给传统的互联网金融营销模式带来了巨大变革。互联网公司可以运用大数据技术对客户在互联网上记录的交易、支付、评价等行为数据信息进行挖掘分析，根据客户的特征、需求和偏好细分客户群体，对客户进行分类管理，针对每一类别的客户定向投放广告和定制产品，从而实现精准营销。例如，支付宝聘请了两家位于硅谷的数据分析实验室从事行为分析，将客户细分成 50 个族群进行研究。亚马逊运用大数据技术对客户的浏览记录、购买行为等进行挖掘分析，进而预测客户的潜在需求。梧桐理财针对能够承担“两万元起投”的中产阶级推出两万元起点的互联网金融理财产品“梧桐宝”，预期年化收益率为 8%～10%。速溶网针对大学生及毕业生推出互联网金融产品“速溶 360”。此外，住金所针对中小微企业的银行贷款周转业务推出了互联网金融产品“安心一过桥贷”。

2. 风险管理

金融创新和金融风险相伴相生。互联网金融在提高金融效率的同时，也带来了一些难以防范的风险。市场风险、信用风险、流动性风险、法律风险、操作风险等都有不同程度的暴露，且交织在一起。例如，P2P 网贷公司倒闭、老板跑路、拆标等的恶意欺诈，资金池、非法集资等违法事件频繁发生。在大数据时代，运用大数据技术能够及时发现风险暴

露，采取措施加以规避和防范。在流动性风险的防范方面，余额宝通过对支付宝的大数据(如客户数量、流量转化率、客户评价等)进行挖掘分析，总结出大量客户申购赎回情况、客户结构、客户行为规律，据此预测出客户下一次申购赎回的时间，从而做出预案以化解流动性风险。在客户流失方面，支付宝根据客户开启和注销账户的数据建立了流失预警模型，进而采取相应的措施争取和留住客户。在系统性风险的防范方面，监管部门通过对大数据的挖掘分析对互联网技能进行实时预警，及时处理突发性事件，防止系统性风险的发生。

3. 信用评价

大数据时代的到来引发了对涉足互联网金融客户信用评价的变革。客户的信用评价不仅包括对评价对象静态信息的分析，还包括动态信息的分析挖掘，同时这也是最重要的。征信机构可以通过大数据技术对客户的注册登记信息(静态信息)以及他们在网络上的购物、支付、投资、生活、公益等数据(动态信息)分析挖掘，形成用户的行为轨迹，通过交叉检验，对客户的真实身份进行识别，进而建立信用评价模型，对客户进行分类，再提供有针对性的服务。例如，阿里巴巴基于淘宝商户的数据，对其电商生态圈内潜在的客户提供纯信用贷款。阿里和腾讯拟推出的“虚拟信用卡”，用户可以实现网上申请，经过对用户交易大数据核查，即可授予一定的信用额度。微众银行通过大数据技术对贷款人的银行储蓄、贷款数据、信用卡数据、社交数据等进行挖掘分析，从而对贷款人进行信用评估，并据此授予贷款人一定的贷款额度。阿里的芝麻信用、腾讯的征信产品、微信的公众号个人信用评分等都是互联网个人征信的开始。

1.5 大数据金融模式

按照大数据服务所处的环节，可以把大数据金融划分为平台金融模式和供应链金融模式。建立在 B2B、B2C 或 C2C 基础上的现代产业通过在平台上凝聚的资金流、物流、信息流组成了以大数据为基础的平台金融，例如阿里金融以及未来可能进入这一领域的电信运营商；建立在传统产业链上下游的企业通过资金流、物流、信息流组成了以大数据为基础的供应链金融，譬如京东金融平台、苏宁易购的供应链金融模式。

1.5.1 平台金融模式

平台金融模式是基于电商平台基础上形成的网上交易信息与网上支付形成的大数据金融，通过云计算和模型数据处理能力而形成的信用或订单融资模式。与传统金融依靠抵押或担保的金融模式相比，不同之处在于：阿里小贷等平台金融模式主要基于对电商平台的交易数据、社交网络的用户交易与交互信息和购物行为习惯等的大数据进行云计算来实时计算得分和分析处理，形成网络商户在电商平台中的累积信用数据，通过电商所构建的网络信用评级体系和金融风险计算模型及风险控制体系，实时向网络商户发放订单贷款或者信用贷款，批量、快速、高效，例如阿里小贷可实现数分钟之内发放贷款。

【案例 1.1】 阿里小贷模式

阿里小贷以“封闭流程+大数据”的方式开展金融服务，凭借电子化系统对贷款人的信用状况进行核定，发放无抵押的信用贷款及应收账款抵押贷款，单笔金额在 5 万元以内，与银行的信贷形成了非常好的互补。阿里金融目前只统计、使用自己的数据，并且会对数据进行真伪性识别、虚假信息判断。阿里金融通过其庞大的云计算能力及数十位优秀建模团队的多种模型，为阿里集团的商户、店主时时计算其信用额度及其应收账款数量，依托电商平台、支付宝和阿里云，实现客户、资金和信息的封闭运行，一方面有效降低了风险因素，同时真正做到了一分钟放贷。京东、苏宁的供应链金融模式则是以电商作为核心企业，以未来收益的现金流作为担保，获得银行授信，为供货商提供贷款。

在阿里小贷业务决策中，数据分析发挥了核心作用。阿里小贷有超过上百个数据模型，覆盖贷前、贷中、贷后管理，反欺诈，市场分析，信用体系，创新研究等板块。其决策系统每天处理的数据量达到 10TB。数据分析用于向公司的管理决策层提供科学客观的分析结果和建议，并对业务流程提出优化改进方案。水文模型就是阿里小贷 2013 年着重搭建的重要数据模型之一。

在信贷风险防范上，阿里小贷微贷技术有完整的风险控制体系。阿里小贷建立了多层次的微贷风险预警和管理体系。具体来看，贷前、贷中以及贷后 3 个环节环环相扣，利用数据采集和模型分析等手段，根据小微企业在阿里巴巴平台上积累的信用及行为数据，可以对企业的还款能力和还款意愿进行较准确的评估。同时结合贷后监控和网络店铺的账号关停机制，可以提高客户违约成本，有效地控制贷款风险。如图 1.12 所示是阿里小贷业务流程。



图 1.12 阿里小贷业务流程

1.5.2　供应链金融模式

供应链金融模式是企业利用自身所处的产业链上下游(原料商、制造商、分销商、零售商)，充分整合供应链资源和客户资源，提供金融服务而形成的金融模式。京东商城、苏宁易购是供应链金融的典型代表。其以电商作为核心企业，以未来收益的现金流作为担保，获得银行授信，为供货商提供贷款。京东商城作为电商企业并不直接开展贷款的发放工作，而是与其他金融机构合作，通过京东商城所累积和掌握的供应链上下游的大数据金融库，来为其他金融机构提供融资信息与技术服务，把京东商城的供应链业务模式与其他金融机构实现无缝连接，共同服务于京东商城的电商平台客户。在供应链金融模式中，电商平台只是作为信息中介提供大数据金融，并不承担融资风险及防范风险等。

【案例 1.2】 京东金融

京东金融于 2012 年开始涉足金融服务，同年，京东金融自主研发产品获得银监会审批，2013 年 12 月推出京保贝。金融业务正在成为京东不可或缺的一部分，而在 2014 年 3 月 7 日，京东低调上线理财产品“小金库”，更证明了京东对于金融领域的野心。

一般企业在与核心企业合作时，既要保证供货，还要承受应收账款周期过长的风险，资金往往成为最大的压力。而这些企业往往因为规模小，资金薄弱，难以得到银行的贷款，资金链断裂成为笼罩在这些企业头上的阴影。京东正是利用用户数据和现有的金融体系，根据每个环链上的业务需求，满足中小微企业的金融需求。

京东做金融有其天然优势，京东有非常优质的上游供应商，还有下游的个人消费者，积累了非常多潜在的金融业务客户。有大数据现成的资源，京东选择金融水到渠成。

在传统的贸易融资中，金融机构只针对单一企业进行信用风险评估并据此做出是否授信的决策，而在供应链金融模式下，银行更加关注的是申贷企业的真实贸易背景、历史信誉状况，而不仅是财务指标。这样，一些因财务指标不达标而难以融资的中小企业，就可以凭借交易真实的单笔业务来获得贷款，满足其资金需求。并且银行通过资金的封闭式运作，确保每笔真实业务发生后的资金回笼，以达到控制贷款风险的目的。

如今，大数据的应用更让京东在这方面如虎添翼。例如，2013 年 12 月推出的京保贝，针对京东上下游合作商提供快速融资的服务，供应商可凭采购、销售、财务等数据快速获得融资。通过大数据，以往需要人工进行的判断、审核等流程可实现自动化审批和风险控制，从供应商申请融资开始，全部由系统实现对放款审核的判断，放款过程全程自动化，因此可以做到 3 分钟融资到账；且无需任何担保和抵押，能有效地提高企业营运资金周转效率。

未来京东金融会覆盖更多的融资服务，而对于产生的数据，包括消费数据、物流数据、供应商财务信息以及金融状况信息，将通过大数据技术进行有效的分析，风险状况也能够实时监控。同时，在了解客户需求的前提下，提供简单融资、快乐融资的融资服务。

(资料来源：数据库频道)

1.6 大数据金融信息安全

21 世纪以来，随着信息技术产业的迅速发展，大数据产业成为新时代背景下继云计算、物联网的发明与广泛应用之后又一大技术产业创新点。金融业通过大数据的应用，催生出基于大数据的客户管理、营销管理、风险管理等应用，商业模式、运营方式、业务模式等不断创新。但在大数据产业呈现爆炸式增长的同时，其大数据信息安全管理水平却呈现非对称发展，所以对现有的信息安全手段提出了更高的要求。特别是大数据技术在金融行业的应用，现在的金融信息化已全面进入信息安全管理阶段，对计算机信息系统有着高度的依赖性，使得金融信息安全面临多方面的威胁，包括大数据集群数据库的数据安全威胁、智能终端的数据安全威胁以及数据虚拟化带来的泄密威胁。大数据时代背景下，高度信息化的金融系统所面临的危险系数更高，必须建立起全方位、多层次、可动态发展的金融安全信息保障体系，以确保金融信息的安全。金融信息安全防范体系可以从这样几个方面完善：建立核心信息区安全防护系统；建立信息交流区安全防护系统；建立内部系统安全防护系统；建立分支节点区安全防护系统；建立管理区安全防护系统。

1.7 大数据应用案例

1.7.1 案例之一：滴滴出行

目前，滴滴已成为整个中国甚至全球发展最快的互联网公司。拥有 3 亿用户，在中国 400 多个城市开展服务，司机超过 1400 万人，1400 万的司机是整个中国所有机动车总量的 10%。每天服务的订单超过 1300 万个，这个订单量让滴滴成为仅次于淘宝的中国第二大互联网交易平台。目前，滴滴平台上每天产生超过 50TB 的数据(相当于 5 万部电影)，超过 90 亿路径规划次数。截至 2015 年 12 月，滴滴出行占据我国网约车市场 46.6%的市场份额，神州专车以 39.9%的比例排名第二，Uber 占 7.2%，排名第三。滴滴、Uber 合并之前，快车市场基本算是两家的二人转；合并之后，占据专车市场 90%以上的份额，算是快车市场的唯一选择。2015 年，滴滴出行平台完成 14.3 亿订单，这相当于在中国平均每个人都使用滴滴打过一次车；累计行驶里程达 128 亿公里，相当于环绕中国行驶 29 万圈，累计行驶时间达 4.9 亿小时，相当于昼夜不歇地行驶 56 000 年。所以滴滴的“数据大脑”对弈的是现实出行的海量数据，通过对每天 24 小时不间断产生的新数据，以及检测这些数据本身产生的二度数据，包括 ETA、路径规划、实际路线、匹配时间等，进行研究、学习，最终实现订单匹配效率的提升，使司机取得更多收入，乘客更加快捷出行。

在业界看来，这巨大订单量背后实则体现的是滴滴出行超强的大数据计算能力。比如，如何将信息推送给更适合区域内的司机、谁优先获得订单，如何给乘客和司机补贴等策略，都要依靠大数据的支持。滴滴根据成交率和应答率来进行智能激励，以此增加用户的叫车意愿，而通过大数据计算，则让订单匹配更加智能，实现了智能派单。例如，以前司机需要开 3 公里才能接到 1 个客人，但现在可能 0.5 公里就能接到客人，在节省时间的

同时，每天的订单成交量也会增加。

1. 滴滴大数据与平台运营管理

1) 供需预测

大数据的神奇之处就在于可以通过搜集到的数据，进行处理分析后，得到规律，然后利用这个规律对未来进行预测。在交通方面，大数据预测的能力极为重要，可以预测什么时间什么地方会拥堵。

大数据预测的关键是有足够多的、高质量的数据。当前滴滴每日峰值订单超过 2000 万单、每日处理数据超过 2000TB，覆盖了交通路况、用户叫车信息、司机驾驶行为、车辆数据等多个维度，它所掌握的真实数据除了可以帮助预测路况外，还能对供需进行预测，供需预测越准确，越能更好地解决供需不平衡问题。

滴滴目前对 15 分钟后供需预测的准确度已经达到 85%，基于这样的准确率，平台可以调度司机满足未来的打车需求，有效降低未来该区域供需不平衡的概率。

2) 路径规划

路径规划和 ETA 两项地图技术是实现智能派单的关键，也将直接影响到司乘双方的使用体验。通过海量历史数据，可以对未来路况做预测，实现 A 点到 B 点的路径规划，它是派单的核心，工程师围绕最低的价格、最高的司机效率和最佳交通系统运行效率来做算法。

ETA 是指预估任意起终点所需的行驶时间，要求精准性。滴滴将机器学习应用到 ETA，这是解决“订单高效匹配”和“司机运力调度”的关键技术。当前滴滴 ETA 可以预测每一单出行的时长以及预估在每一个路口前的等待时长，这项技术可以帮助滴滴在更合适的时间对运力进行更好的调度。ETA/路径规划及其学习系统如图 1.13 所示。



图 1.13 ETA/路径规划及其学习系统

3) 智能派单

滴滴叫车和搜索商品的逻辑不同。网上的商品、资讯等信息都是静态停留在那里，计算方式只是将这个商品、信息挖掘出来；而滴滴的计算则类似于动态打靶，车辆永远在运动当中，要在众多运动的车辆中，给乘客一个最优的选择，不光是距离，还有时间。滴滴研发的基本原则是想办法撮合乘客和司机，满足他们的需求，保证他们的体验。简单点说，就是将订单发送给合适的司机。以滴滴专车业务为例，目前还要用到人为制定的规

则，例如如何将信息推送给最适合区域内的司机、谁优先获得订单等。在数据量较小的情况下，可以基于规则、人的经验来设定算法，但是在数据量更大更丰富的情况下，这样的做法可能和现实存在一定程度的脱节。

这是一个颇为繁杂的过程。除了推荐算法要准确、匹配效率要高、计算要快、推送要及时外，还要在推送订单到这位司机之前，通过对小费、长短途、时间、方向敏感等静态特征和司机与订单之间的位置关系、时间关系等动态特征进行综合分析来预测他对订单感兴趣的程度。智能派单对订单量和司机数进行预测，然后通过大规模分布式计算来实现上述的最优撮合。为了实现这一目的，供需预测、动态调价、路径规划以及服务分的算法技术要一起发挥作用，它们最终为实现最优派单而服务，它们的算法都将结合到智能派单系统中，帮助在动态环境中撮合乘客与司机的交易。

高峰期滴滴平台每分钟接收超过 3 万乘客需求，每 2 秒钟做一次订单匹配，每一次发单背后，滴滴大脑运算次数为百亿次级别。此外，滴滴还可使用大数据技术来预估每个司机的服务分值，包括乘客打分、乘客评价、取消率等因素，并利用算法模型来计算不同服务水平的司机对用户产生的长期影响。

4) 九霄

九霄是滴滴大数据孵化的出行领域智能决策技术产品，能够把错综复杂的时间、空间、业务维度的 N 次元出行领域数据，转化成易于理解的二次元数据，搭建数据理解的桥梁，帮助运营、产品、BI、研发人员发现问题、分析问题、解决问题，产生切实的业务收益。

滴滴将出行领域的数据，进行整理、挖掘、智能聚合，在地图空间和时间轴上进行合理的呈现，使用户能够直观地感知在什么时间、什么地点、各个业务线的什么业务维度(乘客、订单、运力、体验等)，发生了什么，方便深入追踪、探寻业务痛点和原因分析。

例如，通过九霄，对地图上任意区域的供需平衡状况、订单满足情况能够一目了然，并且结合九霄的精细化分析能力，能够细化到某个地理围栏的供需策略，进行围栏级别的运力调度策略配置；(在代驾场景上)基于机器学习进行供需预测，判断哪些区域存在运力缺口，自动化调度司机调节供需平衡。实际上，九霄是凭借科学可视化技术能力、算法能力和高性能架构能力，将数据变为知识，作为决策依据。

此外，给乘客什么样的补贴、给司机什么样的补贴、谁更敏感、多少金额影响更积极，这些策略的背后都是大数据在起作用。

2. 滴滴大数据与城市智慧交通建设

1) 城市道路优化

对大多数用户而言，网约车只是一个打车工具。实际上，网约车背后能做的远远不止这些。每一次的出行背后都是数据的调动和积累，都是对一座城市更为深入的了解。对一座城市而言，尤其是一线城市，网约车有着重大的存在意义。

其一，优化用户出行体验，为用户的出行需求插上互联网的翅膀。

其二，参与优化城市道路交通。习惯网约车后，越来越多的用户正在逐步减少开车次数，一定程度上缓解了道路交通压力和减少了汽车尾气排放。

其三，优化城市道路交通基础规划。

当用户打开滴滴出行 APP 叫车时，毫秒之间，滴滴大脑平均需要 CPU 运算 576 亿次，才能为用户匹配出最优的车辆。这个让人难以置信的数字，深刻反映出滴滴大脑惊人的大数据运算处理能力。推荐上车点“黑科技”上线后，截至 2016 年，滴滴平台上超过 30%的司机和乘客，按照小绿点不需要通话就可以找到对方，司机的通话量平均下降 10%，乘客等候时间平均减少 1 分钟。这些黑科技的背后，其实就是滴滴对一座城市的学习，并且是动态的学习。更好地了解学习城市动态，才能为用户提供更准确的出行服务体验。

2) 滴滴交通云

以武汉为例，滴滴大数据显示，武汉有超过 50%的乘客愿意与别人分享，选择拼车出行。2016 年 1—10 月，武汉有超过 2900 万人次通过拼车和顺风车出行。据测算，一辆充分使用的分享汽车，如果每次行程能够载 2～3 组目的地相近的拼车乘客，每天可减少 20～40 辆私家车上路，因而会大大降低机动车空驶率和上路率。武汉市每天的快车拼车和顺风车出行达 10.2 万人次，如果按私家车平均每辆车每天出行 2 次，每次载客 1.5 人计算，这相当于武汉每天减少 3.4 万辆小汽车出行。

这些庞大的海量数据，都正实时上传到滴滴秘密打造的一朵云——滴滴交通云上。滴滴大脑在这朵云上，根据交通度量体系设定，分析海量数据，让分析结果为乘客、司机、交通主管部门等所有出行参与方都带来价值。滴滴出行正与武汉市交管局共同持续推进武汉“互联网+交通”建设，双方将在路况服务、智能交通云服务等方面进行深度合作。章文嵩介绍，被植入滴滴交通云的城市将发生至少三大变化。

第一，滴滴交通云可以利用智能调度优势帮助改善城市交通拥堵问题。比如，在空间维度，A 到 B 点有很多乘客，滴滴交通云有可能规划不同的行驶道路，让每个路网的车流量均衡。而在时间维度，滴滴交通云可以尝试对早高峰出行的人做精准营销，比如 9 点出行的乘客，如果 8 点出行，补贴 5 元，从而在时间维度上，达到削峰填谷的作用。

第二，滴滴交通云未来还可协助设计智能交通管控方案，提高道路利用率。比如，滴滴交通云可以实现智能信号灯控制，通过数据模型算出整个区域的车流量情况，靠区域的红绿灯协调，让城市各条道路的通行效率更高、更流畅。

第三，滴滴交通云的价值还将体现在，为城市的路网优化提供决策依据。比如，4 个车道，左转弯应该一个还是两个，滴滴交通云都可以给出精准建议。滴滴交通云也会对新建路网做规划建议，比如，应该在哪里建路，或者要不要建一座桥等。

对大众用户而言，很难通过打车去切身体会滴滴大数据的作用，但对于城市交通规划部门而言，交通大数据的采集或许能够针对性解决一些实质问题，尤其是对一些缺乏大型城市交通规划能力的城市而言。

案例小结：滴滴是一个移动互联网产品，依托移动支付，有几乎 100%的支付接入率，滴滴的互联网金融想象空间很大。滴滴将出行连接结构化、数据化，意味着全程可追溯、可评价、可反馈，形成一个促进司机服务不断优化的正向循环生态。滴滴生态里有大量司机资源，滴滴企业平台也是互联网金融好场景，滴滴账户余额也可以变身滴滴版余额宝，只要有海量信任连接的用户，一切都是可以想象的。

1.7.2 案例之二：大数据与美团外卖的精细化运营

美团最初是一个互联网公司，美团的团队之前做校内网(人人网的前身)，后来做饭否网，再后来因为某些原因饭否网被关掉之后，开始转做美团网。2013 年 11 月上线的美团外卖，在两年半的时间内成为中国最大的外卖平台，最近日订单已经突破 400 万份，这个数字放在所有电商交易平台里也能轻松排到前列。而这种高速发展的背后，与大数据技术的支撑和精细化运营是分不开的。美团外卖首先是一个线下商户与线下消费者的线上交易平台，对商户来说，一方面美团可以帮商户解决知名度的问题，可以通过这个平台触达更多的用户。另一方面，由于很多的消费不是发生在线下，商户不需要租用一个大面积、好地段的商铺，可以减少店铺的租金。对消费者来说，可以有更多的选择，并且对现在很多追求生活享受的“宅男”“宅女”或一些生活节奏较快的人也非常方便。

外卖 O2O 和传统的电商存在一些差异。可以简单总结为如下几点。

一是新事物，快速发展。这意味着很多用户对外卖的认知较少，对平台上的新品类缺乏了解，对自身的需求也没有充分意识。平台需要去发现用户的消费意愿，以便对用户的消费进行引导。

二是高频。外卖是个典型的高频 O2O 应用。一方面，消费频次高，用户生命周期相对好判定；另一方面，消费单价较低，用户决策时间短、随意性大。

三是场景驱动。场景是特定的时间、地点和人物的组合下的特定的消费意图。不同的时间、地点，不同类型的用户的消费意图会有差异。例如，白领在写字楼中午的订单一般是工作餐，通常在营养、品质上有一定的要求，且单价不能太高；而到了周末晚上的订单大多是夜宵，追求口味且价格弹性较大。场景辨识越细致，越能了解用户的消费意图，运营效果就越好。

四是用户消费的地理位置相对固定。结合地理位置判断用户的消费意图是外卖的一个特点。

1. 大数据在美团外卖画像技术中的应用

美团外卖经过 3 年的飞速发展，品类已经从单一的外卖扩展到了美食、夜宵、鲜花、商超等多个品类。用户群体也从早期的以学生为主扩展到白领、社区以及商旅，甚至包括在 KTV 等娱乐场所消费的人群。随着供给和消费人群的多样化，如何在供给和用户之间做一个对接，就是用户画像的一个基础工作。所谓千人千面，画像需要刻画不同人群的消费习惯和消费偏好。

1) 外卖产品运营对画像技术的要求

我们大致可以把一个产品的运营分为用户获取和用户拓展两个阶段。在用户获取阶段，用户因为自然原因或一些营销事件(如广告、社交媒体传播)产生对外卖的注意，进而产生了兴趣，并在合适的时机下完成首购，从而成为外卖新客。在这一阶段，运营的重点是提高效率，通过一些个性化的营销和广告手段，吸引到真正有潜在需求的用户，并刺激其转化。在用户完成转化后，接下来的运营重点是拓展用户价值。这里有两个问题。第一个问题是提升用户价值，具体而言就是提升用户的单均价和消费频次，从而提升用户的

LTV(life-time value)。基本手段包括交叉销售(新品类的推荐)、向上销售(优质高价供给的推荐)以及重复购买(优惠、红包刺激重复下单以及优质供给的推荐带来下单频次的提升)。第二个问题是用户的留存，通过提升用户总体体验以及在用户有流失倾向时通过促销和优惠将用户留在外卖平台。所以用户所处的体验阶段不同，运营的侧重点也需要有所不同。而用户画像作为运营的支撑技术，需要提供相应的用户刻画以满足运营需求。如图 1.14 所示为美团用户体验过程。

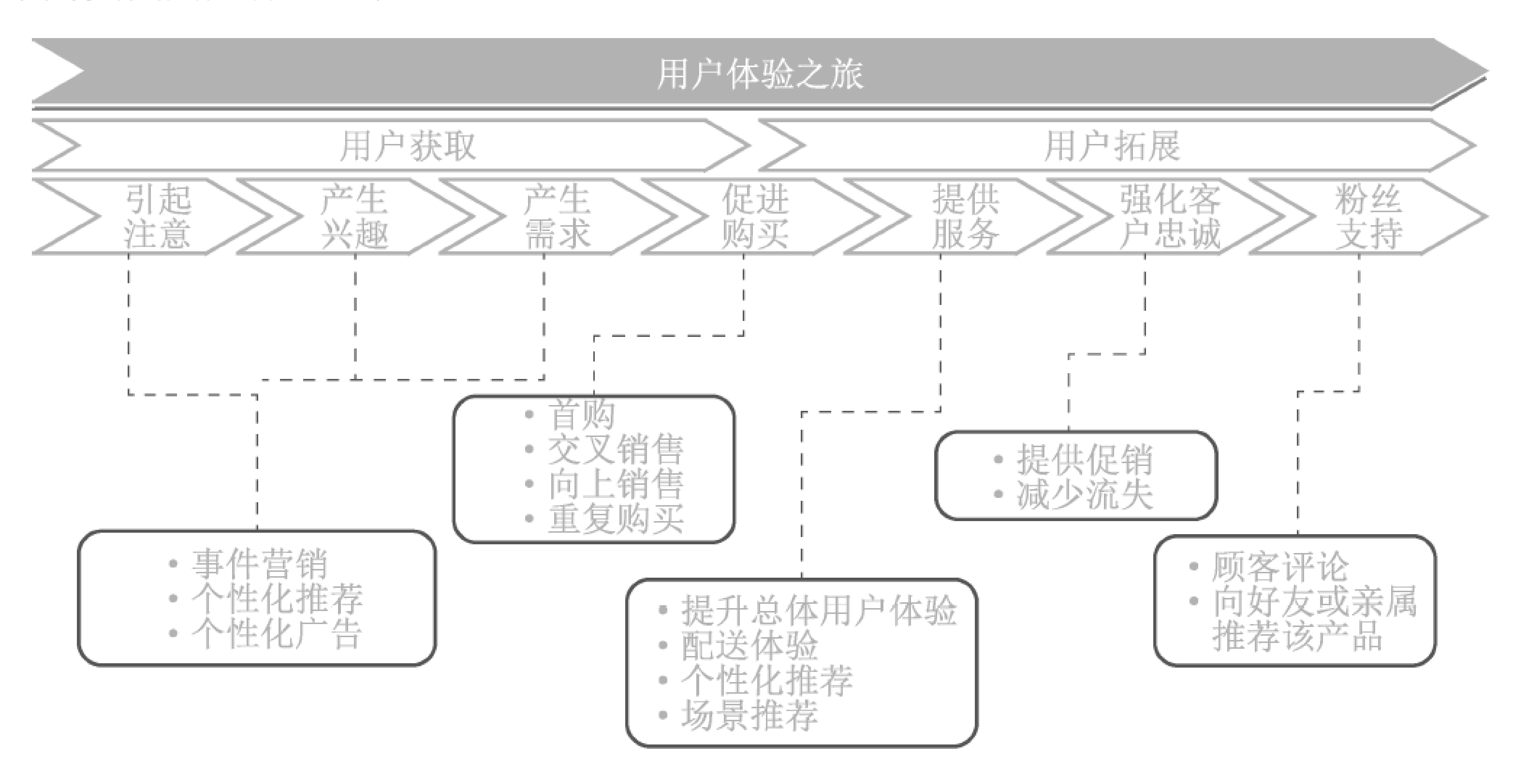


图 1.14 用户体验过程

2) 外卖画像系统架构

画像服务的架构包括：数据源包括基础日志、商家数据和订单数据。数据完成处理后存放在一系列主题表中，再导入 kv 存储，给下游业务端提供在线服务。同时会对整个业务流程实施监控。主要分为两部分，第一部分是对数据处理流程的监控，利用内部自研的数据治理平台，监控每天各主题表产生的时间、数据量以及数据分布是否有异常。第二部分是对服务的监控。目前画像系统支持的下游服务包括广告、排序、运营等系统。如图 1.15 所示为美团画像系统架构。

2. 大数据在美团外卖客户挖掘和预测中的应用

1) 新客运营

新客运营主要需要回答下列 3 个问题。

(1) 新客在哪里?

(2) 新客的偏好如何?

(3) 新客的消费力如何?

回答这 3 个问题是比较困难的，因为相对于老客而言，新客的行为记录非常少或者几乎没有。这就需要通过一些技术手段做出推断。例如，新客的潜在转化概率，受到新客的人口属性(职业、年龄等)、所处地域(需求的因素)、周围人群(同样反映需求)以及是否有充足供给等因素的影响；而对于新客的偏好和消费力，从新客在到店场景下的消费行为可以

做出推测。另外用户的工作和居住地点也能反映他的消费能力。

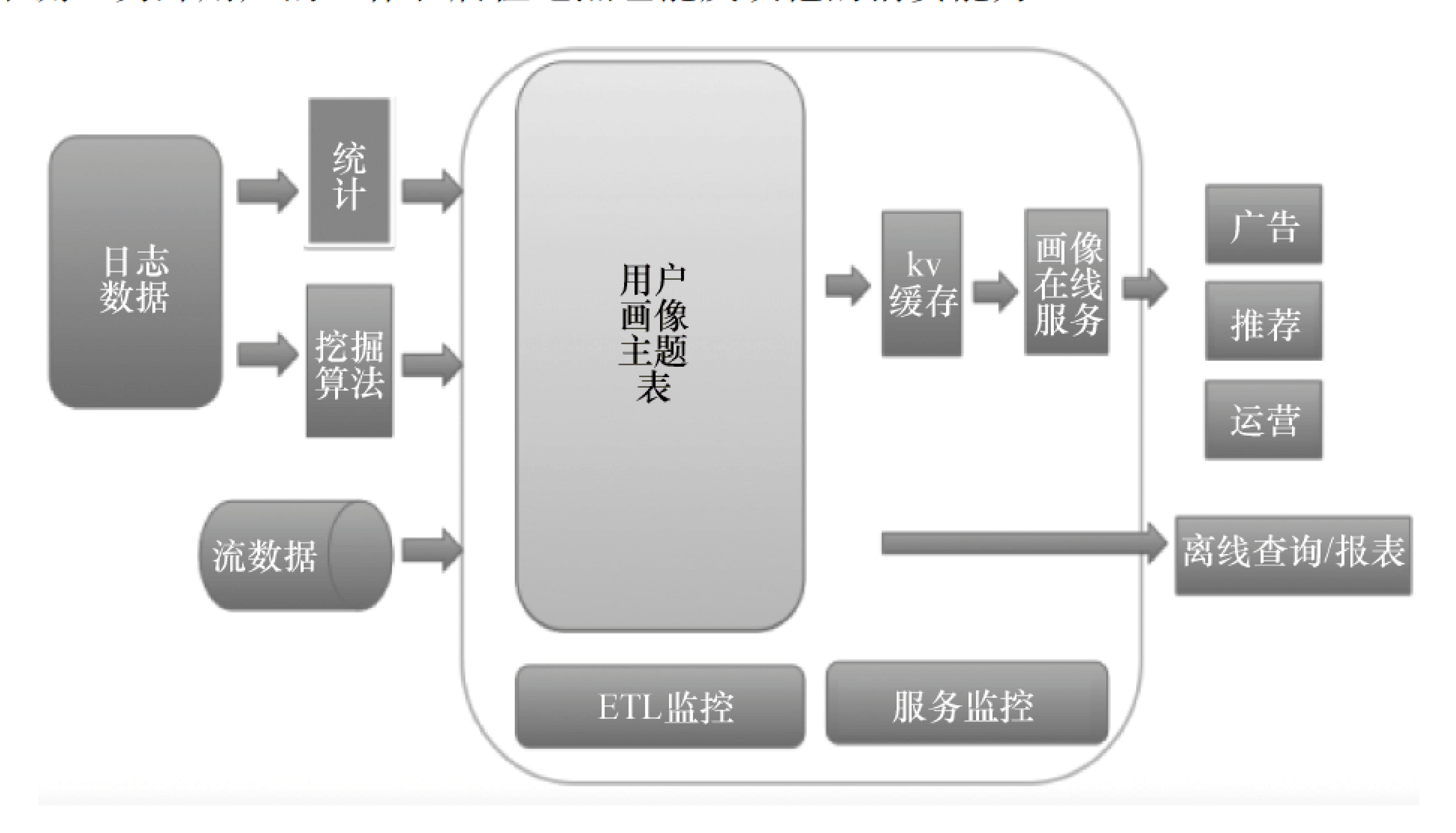


图 1.15 美团画像系统架构

对新客的预测大量依赖于他在到店场景下的行为，而用户的到店行为对于外卖是比较稀疏的，大多数用户是在少数几个类别上有过一些消费行为。这就意味着需要考虑选择什么样的统计量描述，如消费单价，总消费价格，消费品类，等等。然后通过大量的试验来验证特征的显著性。另外，由于数据比较稀疏，需要考虑合适的平滑处理。

美团在做高潜新客挖掘时，融入了多方特征，通过特征的组合最终做出一个效果比较好的预测模型。美团能够找到一些高转化率的用户，其转化率比普通用户高若干倍。通过对高潜用户有针对性的营销，可以极大提高营销效率。

2) 流失预测

新客来了之后，接下来需要把他留在这个平台上，尽量延长生命周期。营销领域关于用户留存的两个基本观点是：获取一个新顾客的成本是维系现有顾客成本的 5 倍；如果将顾客流失率降低 5%，公司利润将增加 25%～85%。

用户流失的原因通常包括：竞争对手的吸引、体验问题和需求变化等。美团借助机器学习的方法，构建用户的描述特征，并借助这些特征来预测用户未来流失的概率。这里有两种做法：第一种是预测用户未来若干天是否会下单这一事件发生的概率。这是典型的概率回归问题，可以选择逻辑回归、决策树等算法拟合给定观测下事件发生的概率。第二种是借助于生存模型，例如 COX-PH 模型，做流失的风险预测。图 1.16 左边是概率回归的模型，用户未来 T 天内是否有下单作为类别标记 y，然后估计在观察到特征 X 的情况下 y 的后验概率 $P(y|X)$。右边是用 COX 模型的例子，我们会根据用户在未来 T 天是否下单给样本一个类别，即观测时长记为 T。假设用户的下单的距今时长 $t<T$，将 t 作为生存时长 t'；否则将生存时长 t'记为 T。这样一个样本由三部分构成：样本的类别(flag)、生存时长(t')以及特征列表。通过生存模型虽然无法显式得到 $P(t'|X)$的概率，但其协变量部分实际反映了用户流失的风险大小。

生存模型(见图 1.17)中，$\beta^T x$ 反映了用户流失的风险，同时也和用户下次订单的时间间隔成正相关。在箱线图中，横轴为$\beta^T x$，纵轴为用户下单时间的间隔。

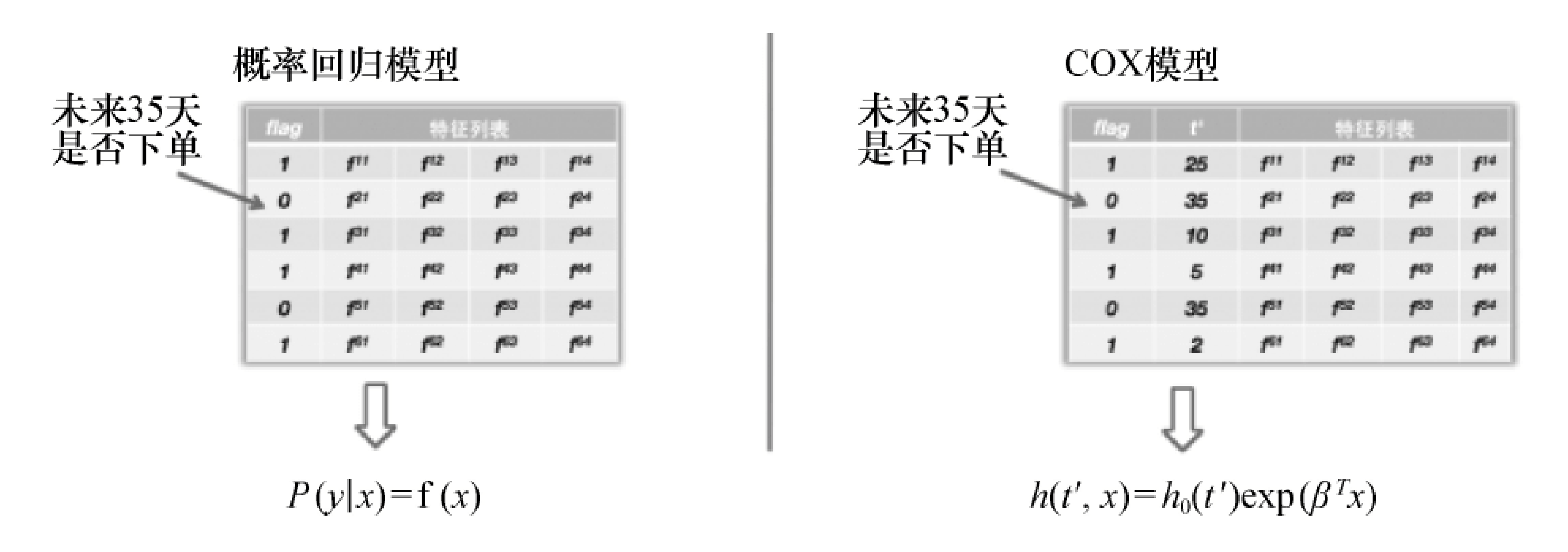


图 1.16　流失预测模型

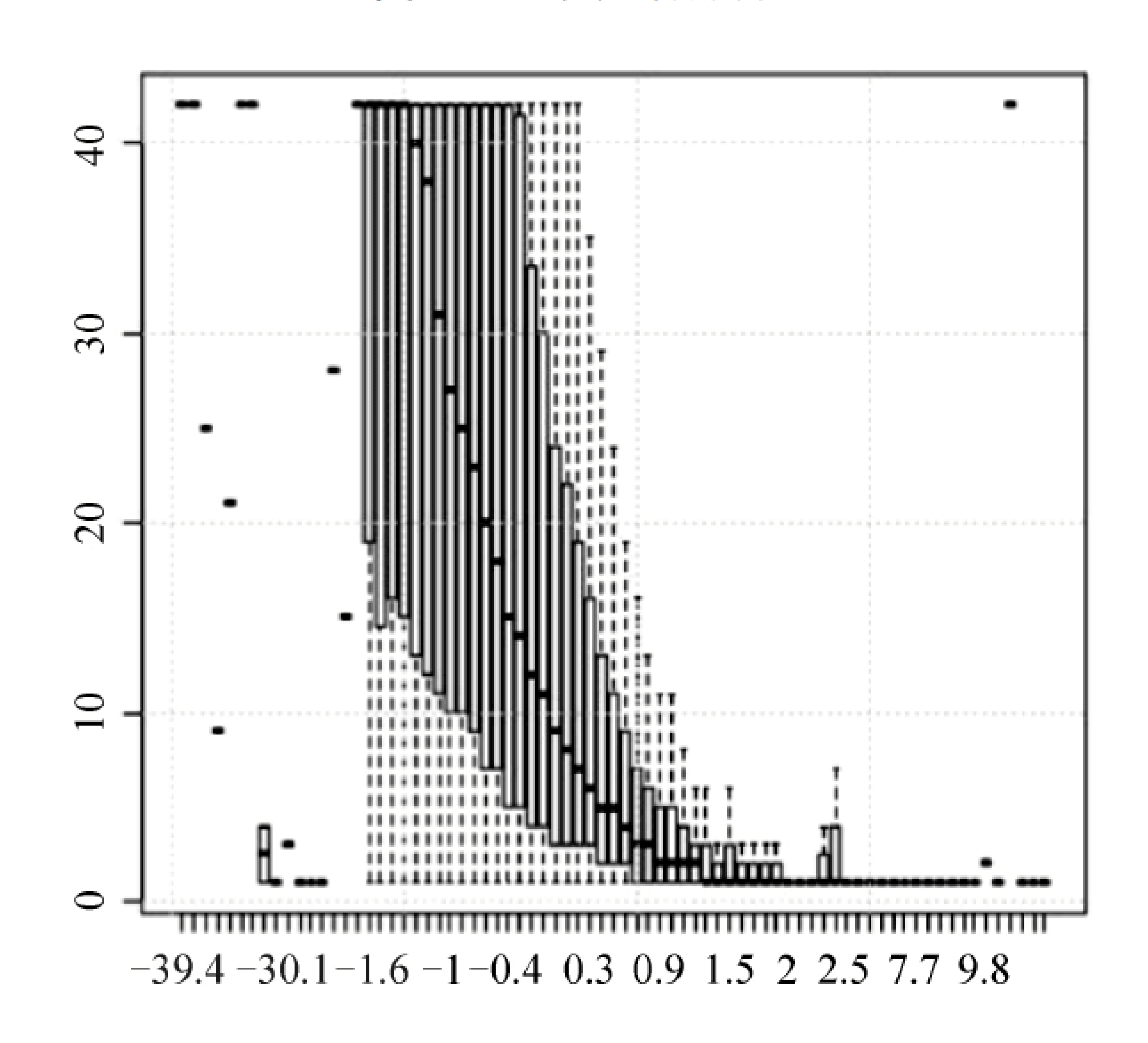


图 1.17　生存模型

美团做了 COX 模型和概率回归模型的对比。在预测用户××天内是否会下单上面，两者有相近的性能。美团外卖通过使用用户流失预警模型，显著降低了用户留存的运营成本。

3. 大数据在美团外卖用户补贴中的应用

美团外卖吸引顾客的一个方式是用户补贴。用户补贴对于平台而言是一笔巨大的运营成本。但是在很多情况下，用户补贴是很有必要的。平台都希望吸引更多的新用户以及留住老客户，这是业务发展的重中之重。那么，怎样进行用户补贴才能有助于平台吸引客户，之后源源不断地在平台上消费，这就需要大数据分析作为支撑。

首先，以客户留存率和自动转化意愿将用户群体划分为四个象限(见图 1.18)。第一个维度是客户留存率。在互联网行业中，用户在某段时间内开始使用应用，经过一段时间

后，仍然继续使用应用的被认作是留存客户；这部分用户占当时新增用户的比例即是客户留存率。需要关注的是，对于刚开始使用美团外卖的新用户，有多大可能性会一直留在平台，在平台不给补贴的情况下，还会不会继续留在平台。有的用户在只有给红包的情况下才会留下，这就是留存率低的情况。另一个维度是用户的自动转化意愿。平台上每天都会有很多新客户，有的新客户没使用红包就开始使用平台的服务，有的客户只有给他发红包才能完成转化，但是转化之后会成为平台的忠实客户，有的客户一旦没有红包就会自动流失。那么，对于自动转化意愿高的用户，不给红包也愿意使用，如果能识别出来这类客户，就可以不给补贴。还有一些只有收到红包才使用，之后又会流失，如图中第三象限，这部分用户最好不发补贴。平台需要识别给了红包就可以一直留在平台，即使以后没有红包也会一直保持消费忠诚的客户。要知道客户属于哪个象限的，就需要用大量数据做挖掘。

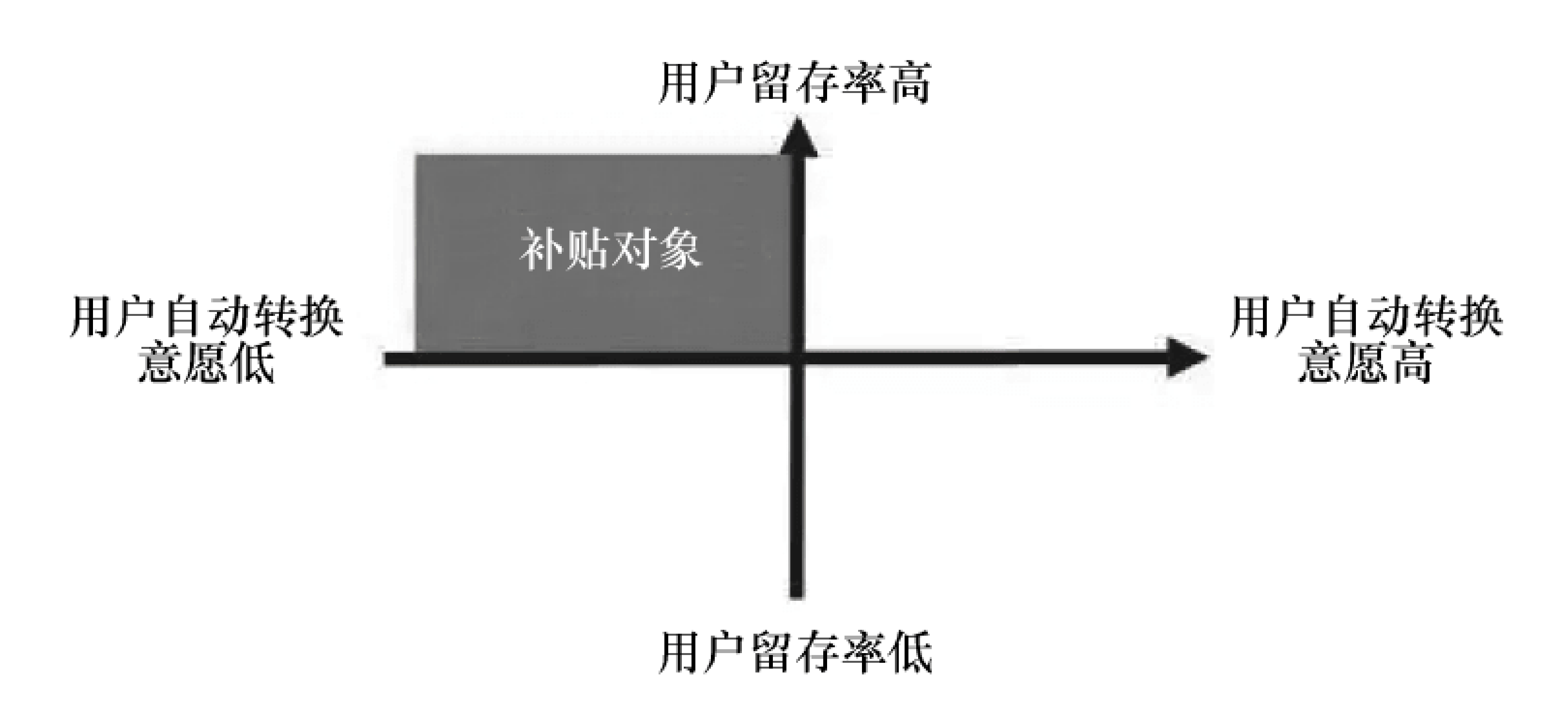


图 1.18　用户群体划分

为了识别不同类型的用户，需要做用户画像(见图 1.19)，首先通过各种渠道了解用户的年龄、婚姻状况、收入水平、消费习惯、常住地等，了解这些信息后，需要做一次用户模型的训练，即把各种参数跟四个象限做一次训练，就知道给什么样的用户推红包。以下分析如何使用用户画像来寻找补贴对象。

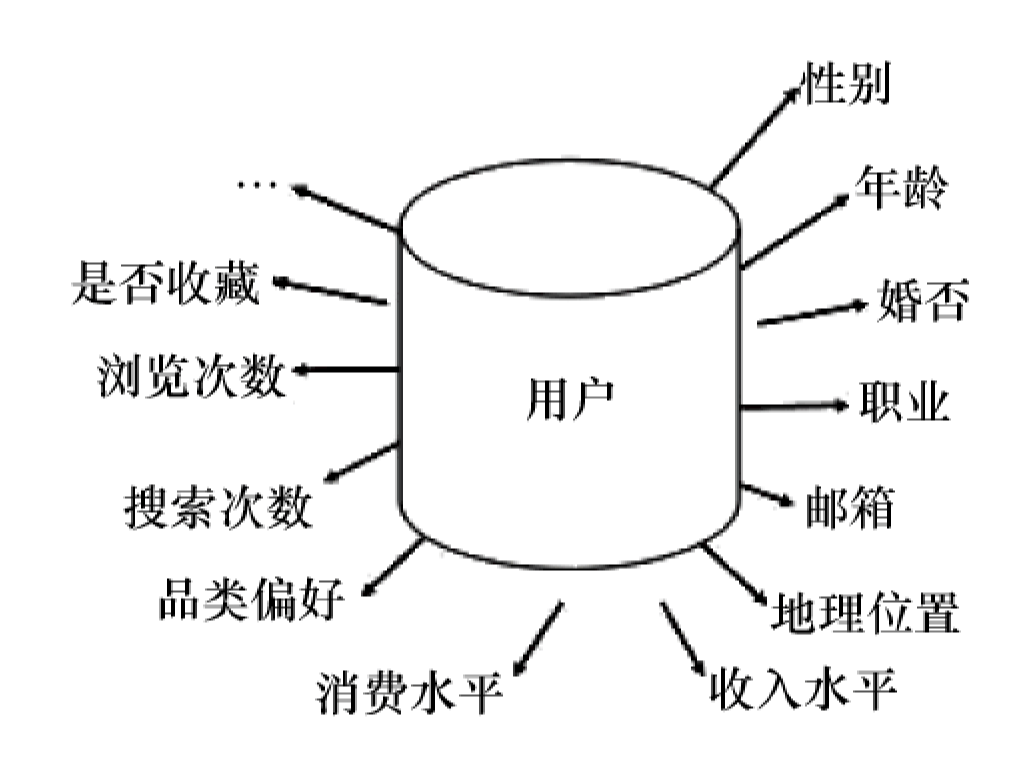


图 1.19　用户画像要素

图 1.20 是数据挖掘的流程示意图。美团外卖也是一个大数据公司，每天有大量用户浏览网页、大量用户在网站上购买、用户的位置属性和手机型号这些数据也都可以获得，另

外还有用户自己填写的性别、年龄等，这些都可以用来做用户数据的挖掘。有了这些数据之后，通过数据挖掘的算法，对用户进行深入的挖掘之后，再去完善用户的各种画像，就可以得到一个模型。然后对于任何一个特定的用户，将其浏览购买历史的信息输入模型中，就可以知道这个用户很多具体的特征。例如，用户的年龄为 25 到 30，可能是一个刚毕业没多久的白领。例如，用户是一个夜猫子，那么平台就在晚上向其推荐一些夜宵。再比如，单身理工男，他可能就不太喜欢甜食、水果。这就是平台通过大数据挖掘得到的很多用户画像的信息，之后将这些信息用于平台用户营销、用户补贴等方面，可以快速提高平台识别和分析用户的精准度。如果理想情况下，把原来随便撒红包的形式，限制在一个象限，平台资金的使用效率就会更高，这就是大数据在用户补贴方面的使用。

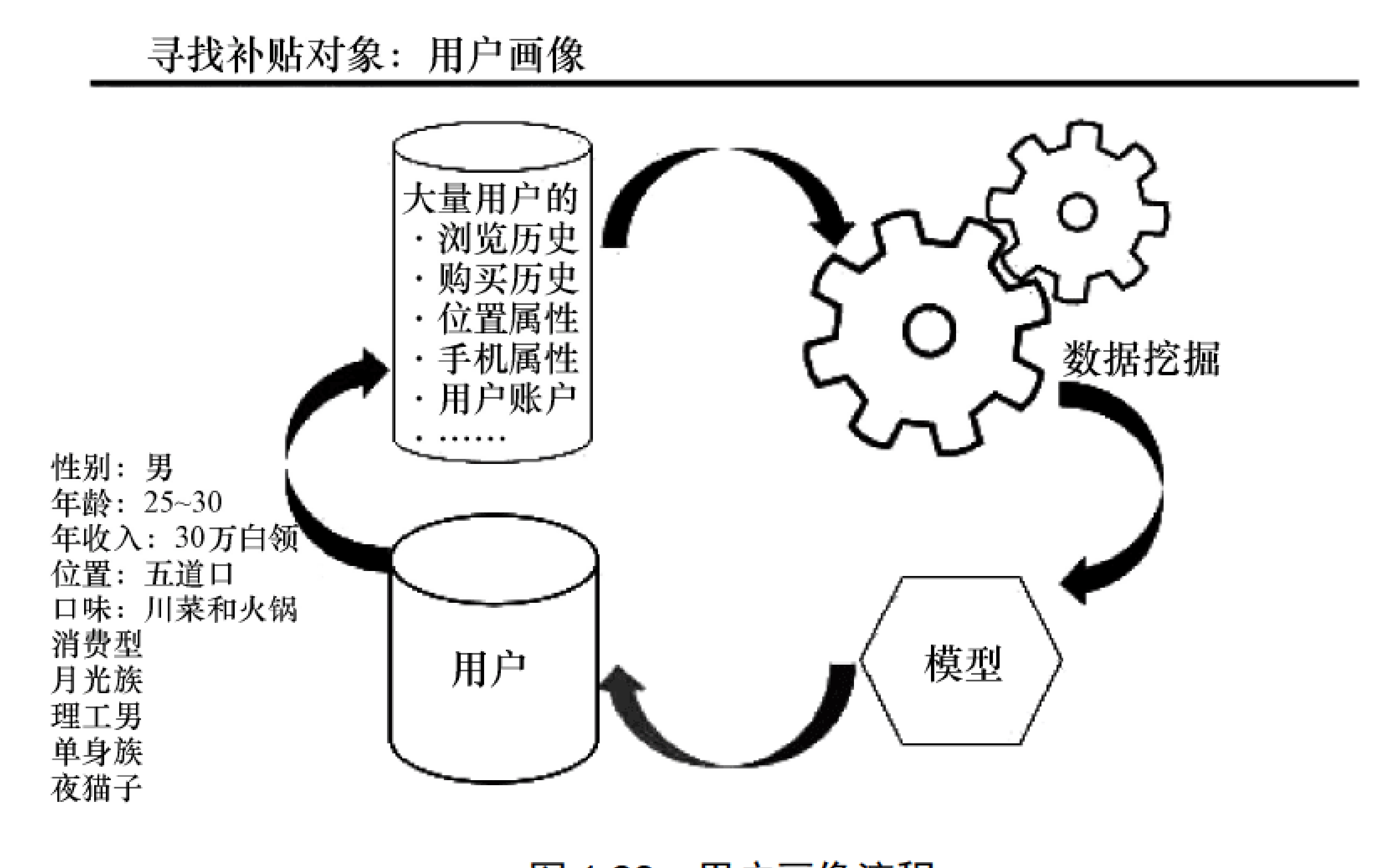


图 1.20 用户画像流程

4. 大数据在美团外卖供应链中的应用

“外卖 O2O”是一种特殊的 O2O 形态，多了一个配送和调度的部分。其他如信息发布、用户信息的搜索以及支付和以往都一样，只是履约部分由消费者到商家店里消费改成配送上门。配送团队首先报告给平台，每一个配送员的位置在什么地方，平台根据这个定位和订单的信息发起调度，告诉配送员去取哪个订单，然后这个配送员就可以去商户那边把订单取回来，送给消费者，这是外卖 O2O 的模式。

1) O2O 闭环

外卖 O2O 模式的真正价值在于通过线上工具与云服务器、CRM、餐饮管理系统的信息化无缝闭合回路，依靠云计算功能处理、转化、应用大数据。所以，只有无缝整合线上线下资源，形成 O2O 闭环才是企业能否踏入大数据时代的关键。外卖 O2O 闭环的难点和痛点在于，如何收集线下消费者体验的反馈信息，并将线下用户引到线上交流，进行线上体验。有的商家以为自己促使消费者完成线上的支付就是 O2O 闭环，这是肤浅的想法。把消费者从线下送到线上，这个线上不仅仅是支付，而是要形成线上的消费者与消费者、消费者与商家之间的互动。

互动需要一个能够容纳消费者和商家的平台。现在看来，微信平台是一个不错的选

择。可以在微信上建立一个公众平台，与消费者进行互动交流。互动的话题可以是让消费者自己来设计喜欢的菜单，让消费者自己来设计喜欢的菜品。一旦得到采纳，消费者可以获得奖品，或者是此菜品永远对该设计者免费。这样，除了在平台上完成订餐、点菜、支付等功能外，还能根据消费者的消费行为有针对性地进行推广和促销。更重要的是，能充分发挥粉丝经济的作用，让粉丝参与到服务改进、菜品改进中来，提高顾客的满意度并提升商家的销量和形象。

自此，把消费者从线下引到线上的途径就达成了，这样才是一个真正的 O2O 闭环(见图 1.21)。从线上到线下，再从线下到线上，消费者在被引导，数据和信息才会流入商家的口袋中。

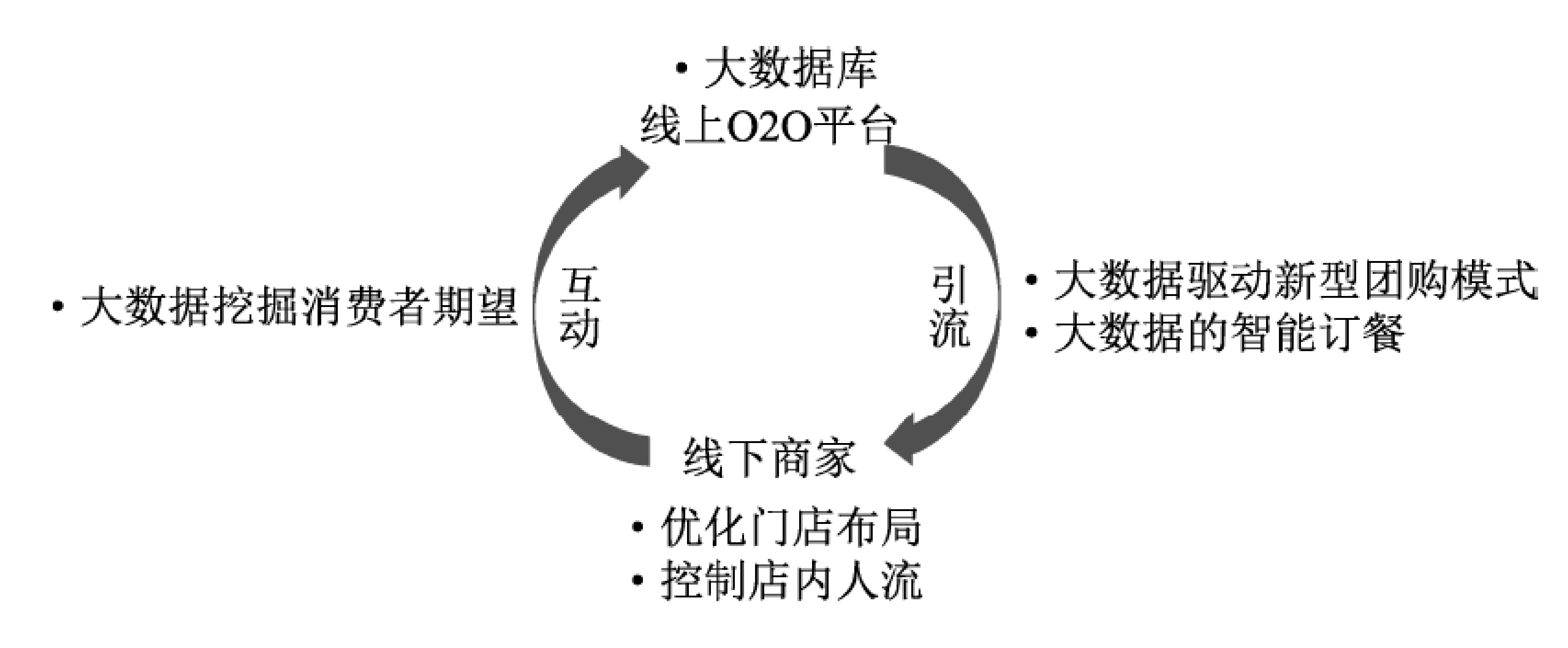


图 1.21　大数据库 O2O 闭环

2)　配送的智能调度系统

如图 1.22 所示为配送调度的示意图，黑色的袋鼠是配送员(骑手)。一般而言，平台没法预测用户下单的位置。每个用户下单的频率不一样，下单的餐馆也不一样。这些骑手站位要怎么站位，这么多订单要让哪个骑手取哪个订单，用什么顺序送这个订单，这些在极大程度上影响了骑手的能效。如图 1.23 所示为配送要考虑的因素。

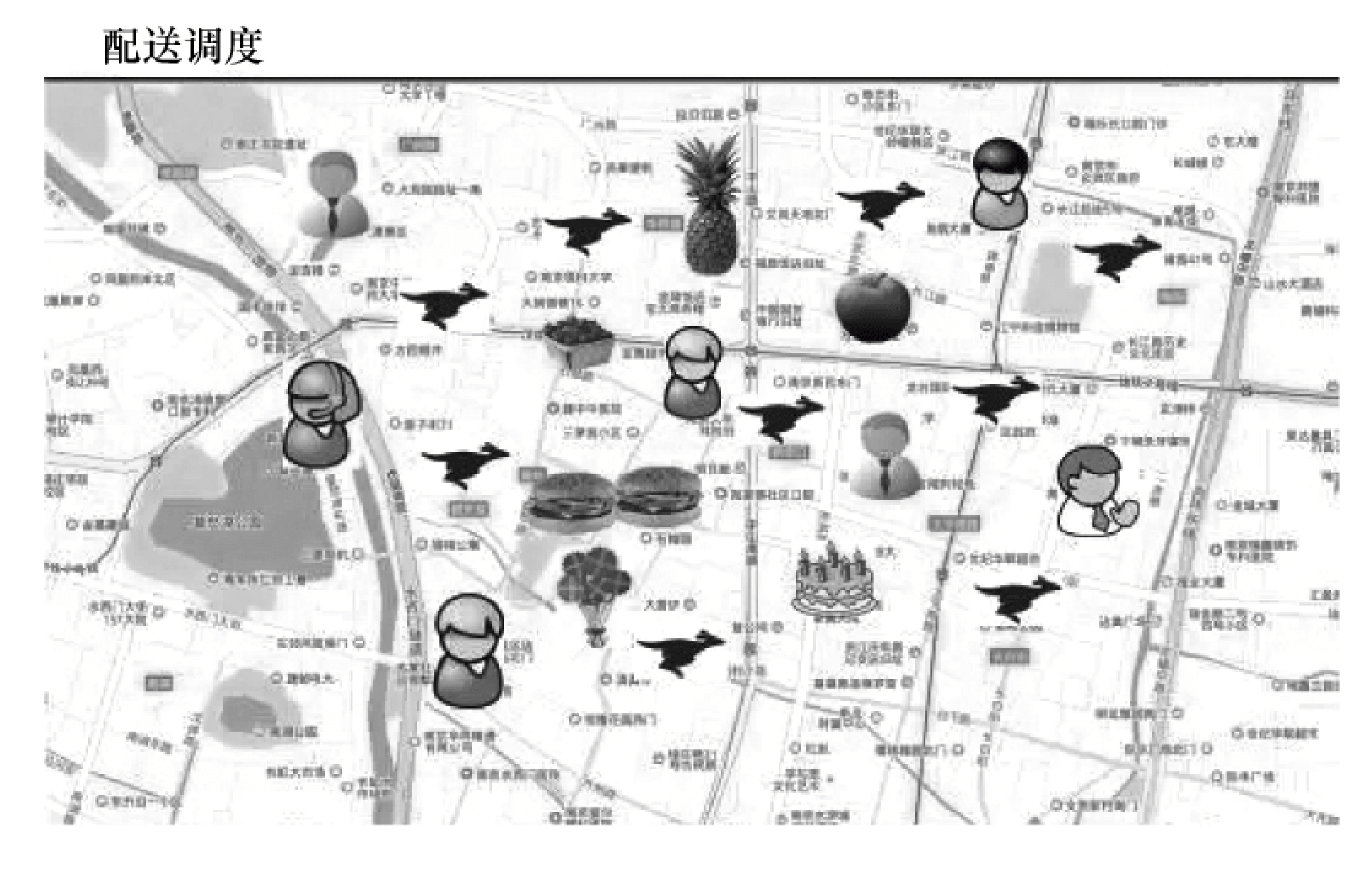


图 1.22　美团外卖配送调度图

外卖配送问题

配送要解决的核心问题

将订单分配至骑手

提供任务执行路线

要考虑的因素：
1. 运单类型、下单时间、期望送达时间、商户地址、用户地址、预计出餐时间、调度轮次。
2. 骑手数量、正在执行的运单状况、当前经纬度、最近一次上报坐标时间、当天已完成的订单量、骑手个性化因素(速度、评分、接单偏好等)。
3. 环境：骑手/商户/用户之间的距离、天气/节假日、当前总体负载、未来负载预测值、调度类型(众包/自建/加盟)

图 1.23 外卖配送问题

如果考虑一个数学模型(见图 1.24)，目标函数就会优化很多的变量，比如说每一单平均的行驶距离，因为平台根据这个给配送员付工资；配送时间，这个决定了用户的体验；运单的准时率；如果订单很多，骑手有限，实在送不过来，希望最坏的体验也不要太坏；还有骑手的满意度，在平台技术团队看来，骑手也是平台的客户，所以预期在分派订单的时候骑手也是满意的。

配送调度问题的数学模型

类别	内容
目标函数	最小化单均行驶距离
	最小化单均配送时间
	最大化运单准时率
	最小化最坏配送时长
	最大化骑手派单满意度
决策变量	各运单分配的骑手
	骑手建议行驶路线
约束条件	行驶速度约束
	最大单量约束
	出餐时间约束
	交付时间约束
	骑手可用约束等

- 组合优化难题
- 非线性、不确定
- 实时性要求高
- 强NP-Hard难题

研究现状：
- 学术界没有高效的求解方法
- 工业界在进行不同的方法探索

图 1.24 配送调度问题的数学模型

如图 1.25 所示为美团外卖现在使用的一套系统。平台同样会用很多的特征数据来挖掘，包括给骑手画像，即骑手的骑程速度、送多少，他的箱子能送什么，骑什么车送外卖，等等。给商家画像，即这个商家一天最多出多少单，可以做多少盒外卖，商家从接到订单到准备好订单平均大约几分钟，等等。还有各种配送指标，包括统计指标、骑手上报的各种路径，然后都会放在其中进行挖掘。挖掘之后会结合所做的一个数学模型在一个仿真平台上运行，运行之后发现这个算法可以，就放到一个实时的海量计算平台上给实时的这些订单做实时调度。调度算法是基于一个多目标的运筹优化的数学模型。在做调度的过程中，会同时监控每一个订单，监控实际配送时间跟预期的配送时间是否相同，骑手实际

走什么样的路径，跟系统预期的最佳路径是否相同，还包括一些骑手给平台反馈一些不好的调度方案等。平台把这些信息都收集回来，再输入到回放平台，找到那个时间点再回放一下，这样可以指导这个算法有什么地方可以优化，或者需要添加哪些新的数据参数。

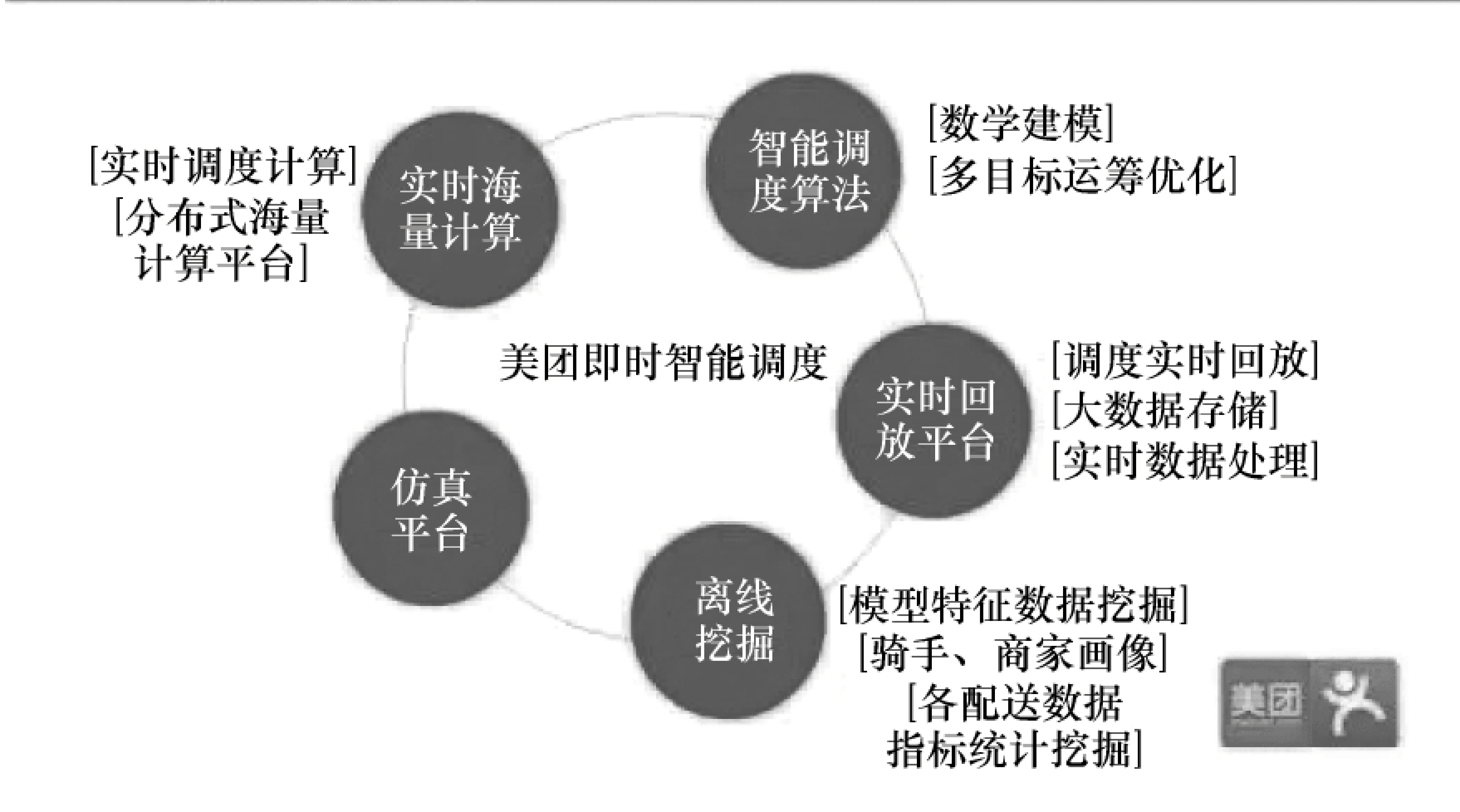


图 1.25　美团外卖配送的智能调度系统

这个系统数据量特别大，同时包括骑手的数据，十几万骑手，每十秒上报一个位置，所以有特别多的点。再加上给全国很多城市做配送，美团专门跟气象部门合作，购买了每个城市的实时气象数据，然后系统会根据气象数据知道城市的雨雪情况，根据这个天气情况做调度，在这里需要输入的数据特别多。基于这种情况，平台使用大数据技术对配送的调度方案进行不断的探索和优化。

5. 大数据在扼制恶意刷单套现中的应用

有用户补贴往往会带来刷单作弊，一般有利益的地方就容易产生这种情况。刷单就是一种行业毒瘤，简单来说就是一些作弊的订单和作弊的交易。刷单者有很多不同的目的，比如说用大量虚假的订单套取补贴的利率、套现，以及制造一些虚假的订单量。例如，一个店铺入驻美团外卖之后，本来一个月卖三单五单，通过大量的虚假订单，一个月一下子能有 1000 单的销量。因为美团外卖网站按照销量对商家进行排名，通过刷单把月销量刷上去，店铺排名就在上面，就会带来更多的流量，这就是制造虚假订单。还有一部分是利用虚假订单写一些虚假的评论来误导用户。跟补贴相关的，主要是补贴套现。

与刷单行为做斗争对 O2O 行业的发展是非常重要的一件事情。美团最初的做法，是在网页上面放了一个链接，让大家举报刷单，运用群众的力量抑制刷单。然而这种方法没有什么用，因为真正刷单的人，如在家里用“猫池”刷单的人，别人是无法知道的。事实证明，通过这种方式接到的举报很少，而接到的举报经查，往往是商家的一些竞争对手为了打击对方的虚假举报。在这种情况下，只有靠大数据的技术手段才能抑制刷单，如图 1.26 所示。

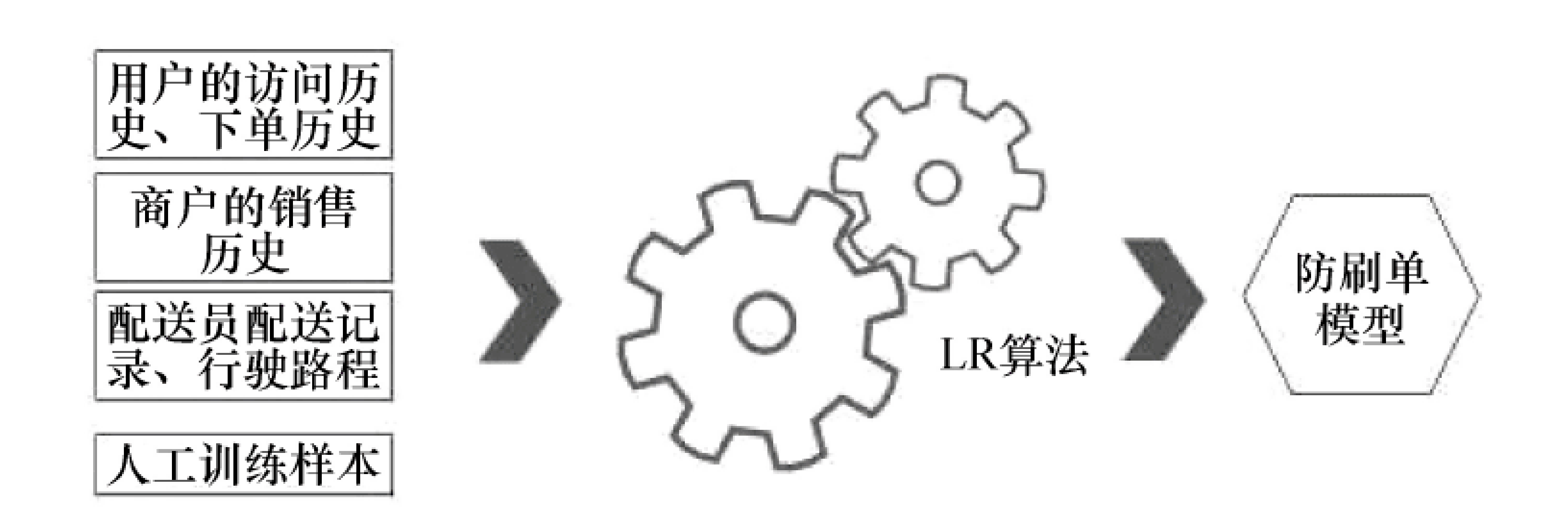


图 1.26　美团外卖防刷单模型

图 1.26 只是一个简单的示意图，美团外卖平台会收集大量的数据，包括每个月用户所有的下单历史、浏览历史，商家的销售历史，例如一个商家销售 1000 单，这 1000 单到底是卖给了一个用户还是卖给很多个用户，这里面可以找到很多规律。美团现在每个订单都是由配送员配送的，每个配送员的 APP 要每隔 10 秒汇报一次他的地理位置，所以平台有所有配送员的路径记录。同时有一支运营团队去人工分析，哪些订单是刷单的，有哪些特征。利用这些人工的样本作为种子，加上上面的一些数据，同样的把它们输入到整个系统的大数据库里面。最后有一个防刷单模型，基于这个刷单模型就可以判断某个用户刷单的可能性和商家刷单的可能性。对可能性比较高的，平台运营人员会介入并分析，如果是真的就予以比较严厉的惩罚，情节较轻就追回刷单的赃款，对于情节较重的行为将以欺诈的罪名，上报相关法律机构。

案例小结：互联网企业具有高速的发展潜力，互联网企业的竞争正在从资本驱动，慢慢走向技术驱动、数据驱动。对这些企业而言，整体上由原来的补贴导向的竞争，慢慢变成靠公司整体的运营治理、技术和数据导向的竞争。

本章总结

- 大数据是指在一定时间范围内无法用传统数据库软件进行采集、存储、管理和分析的数据集或数据群，需要通过新的处理模式才能体现出的具有高效率的、高价值的、海量的、多样化的信息资产。大数据具有大体量、多样性、时效性、准确性、价值性这 5 个特征。
- 小数据是以个体为中心，需要新的应用方式才能体现出的具有高价值的、个体的、高效率的、个性化的信息资产。大数据和小数据有着本质的区别，虽然两者都是以创造数据价值为目的，但是收集目的、数据结构、生命周期、分析方法及分析重点方面都存在着不同的定位。
- 大数据的分类形式众多。按照大数据的结构特征，可以将大数据分为结构化数据、非结构化数据和半结构化数据。按照大数据的获取处理方式，可以将大数据分为批处理数据和流式计算数据。按照大数据的处理响应性能，可以将大数据分

为实时数据、非实时数据和准实时数据；按照大数据的关系，可以将大数据分为简单关系数据和复杂关系数据。

- 大数据的处理流程归纳为：首先利用多种轻型数据库收集海量数据，对不同来源的数据进行预处理后，整合存储到大型数据库中，然后根据企业或个人目的和需求，运用合适的数据挖掘技术提取有益的知识，最后利用恰当的方式将结果展现给终端用户。具体包括数据采集、数据预处理、数据存储、数据挖掘及数据解释这 5 个步骤。
- 大数据金融是指运用大数据技术和大数据平台开展金融活动和金融服务，对金融行业积累的大数据以及外部数据进行云计算等信息化处理，结合传统金融，开展资金融通、创新金融服务。
- 大数据金融与传统金融相比，存在如下几个方面的特点：呈现方式网络化；风险有所调整；信用不对称性大大降低；金融业务效率提高；金融企业服务边界扩大；产品是可控的、可接受的；惠普金融。相对于传统金融，大数据有着无可比拟的优势：放贷快捷，精准营销，个性化服务；客户群体大，运营成本低；科学决策，有效风控。
- 大数据给传统的金融业、征信业和新兴的互联网金融行业带来了较大的变革。与此同时，还带来了较大的金融信息安全问题和监管挑战。因此，我们在享受大数据带来的价值的同时，还应该建立起完善的安全防范体系，以确保金融数据信息的安全。
- 按照大数据服务所处的环节，可以把大数据金融划分为平台金融模式和供应链金融模式。平台金融模式是基于电商平台基础上形成的网上交易信息与网上支付形成的大数据金融，通过云计算和模型数据处理能力而形成信用或订单融资模式。供应链金融模式是企业利用自身所处的产业链上下游，充分整合供应链资源和客户资源，提供金融服务而形成的金融模式。
- 在大数据背景下，金融信息安全面临多方面的威胁，包括大数据集群数据库的数据安全威胁、智能终端的数据安全威胁以及数据虚拟化带来的泄密威胁。

本章作业

1. 大数据的内涵是什么？与小数据有什么区别？大数据有哪些特征？
2. 大数据与传统数据有哪些区别？
3. 大数据的价值体现在哪些方面？
4. 大数据在金融业中有哪些应用？
5. 大数据金融的内涵和特点是什么？
6. 与传统金融相比，大数据金融有哪些优势？
7. 大数据给银行业、保险业、证券业、征信业分别带来了哪些大变革？
8. 大数据金融在互联网金融领域中有哪些应用？
9. 大数据金融信息存在哪些安全问题？如何解决？

第 2 章

大数据相关技术

本章目标

- 掌握大数据处理流程：数据采集、预处理、存储、挖掘和解释
- 掌握大数据的 3 种来源：核心数据、外围数据、常规渠道数据
- 掌握大数据的主要架构
- 掌握数据挖掘常用方法

本章简介

本章从大数据处理流程、数据来源、大数据生态圈及主要架构、数据挖掘的主要方法几个方面来介绍大数据的相关技术。

2.1 大数据处理流程

大数据的处理流程归纳为：首先利用多种轻型数据库收集海量数据，对不同来源的数据进行预处理后，整合存储到大型数据库中；然后根据企业或个人目的和需求，运用合适的数据挖掘技术提取有益的知识；最后利用恰当的方式将结果展现给终端用户。具体包括：数据采集、数据预处理、数据存储、数据挖掘及数据解释这 5 个步骤，如图 2.1 所示。

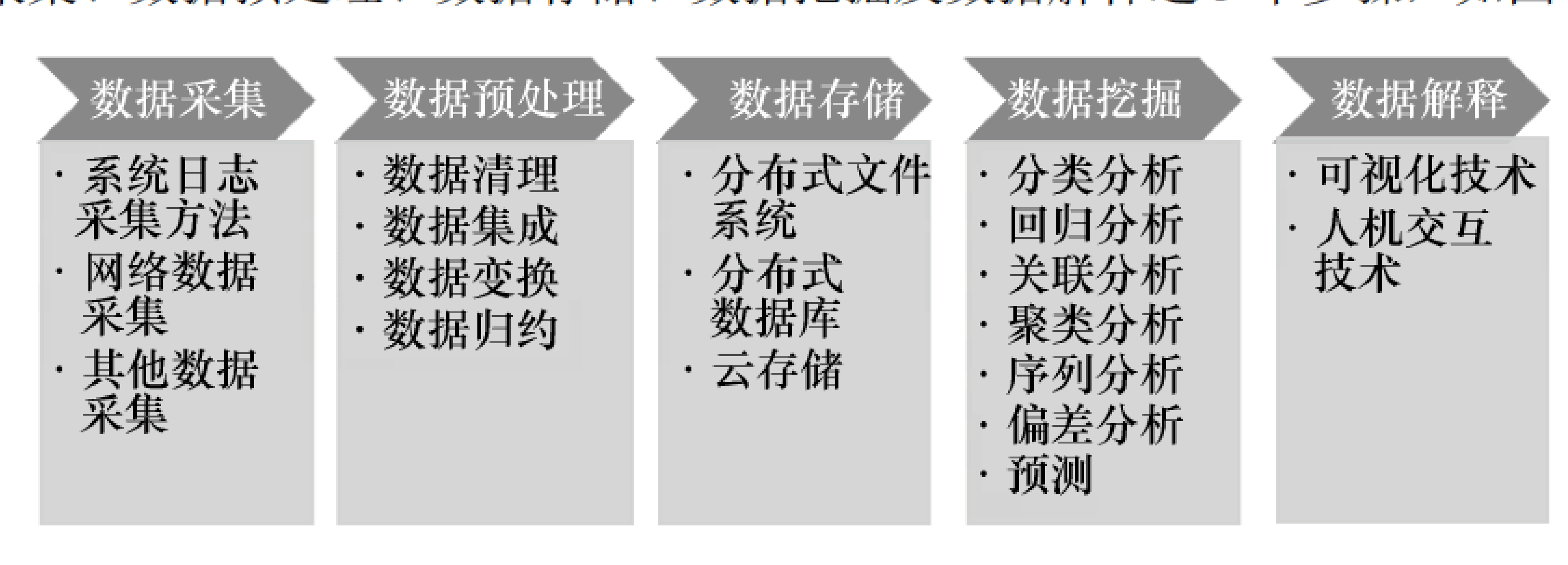


图 2.1　大数据的处理流程

2.1.1　数据采集

大数据的采集是大数据处理过程中的第一步，它是数据分析和挖掘的基础。大数据的采集是指在确定用户目标的基础上，对该范围内的所有结构化、半结构化、非结构化数据进行采集的过程。采集的数据大部分是瞬时值，还包括某时段内的特征值。大数据的主要来源有商业数据、互联网数据、传感器数据。针对不同来源的数据，具有不同的采集方法。主要的大数据采集方法有系统日志采集方法、网络数据采集方法、其他数据采集方法。

1. 系统日志采集方法

大多数互联网企业都有自己的海量数据采集工具，常用于系统日志采集，如 Scribe、Flume、Chukwa、Kafka 等。Scribe 是 Facebook 开源的日志收集系统，能够从各种日志源收集日志，存储到一个中央存储系统中，以便于进行集中统计分析和处理；Chukwa 属于 Hadoop 系列产品，是一个大型的分布式系统监测数据的收集系统，提供了很多模块以支持 Hadoop 集群分析；Flume 是 cloudera 的开源日志系统，能够有效地收集汇总和移动大量的实时日志数据。这些工具均采用分布式架构，能满足每秒数百 MB 的日志数据采集和传输需求。

2. 网络数据采集方法

网络数据采集是指利用互联网搜索引擎技术从网站抓取数据信息。目前，网络数据的采集基本上是利用垂直搜索引擎技术的网络爬虫或数据采集机器人、分词系统、任务与索引系统等技术进行综合运用而完成。该方法可以将非结构化数据从网页中抽取出来，将其存储为统一的本地数据文件，并以结构化的方式存储。它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。除了网络中包含的内容之外，对于网络流量的采

集可以使用 DPI 或 DFI 等带宽管理技术进行处理。

3. 其他数据采集方法

对于企业生产经营数据或学科研究数据等保密性要求较高的数据，可以通过与企业或研究机构合作，使用特定系统接口等相关方式采集数据。

在大数据的采集过程中，同一网站同一时间可能会有很多用户访问和操作。例如，火车票售票网站和淘宝，它们并发的访问量在峰值时超过了上百万，并发数十分高。因此，需要在采集端部署大量数据库才能支撑。

2.1.2 数据预处理

由于第一步收集得到的数据是原始数据，存在着不完整、不一致的问题，无法直接存储到数据库中进行数据挖掘。因此，在将来自前端的数据导入一个集中的大型数据库或者分布式存储集群前，需要对大数据进行预处理，这样不但能够节约大量的空间和时间，还能得到更好的数据挖掘结果。大数据预处理包括对数据进行清理、集成、变换和归约 4 个过程。

1. 数据清理

数据清理是数据准备过程中最乏味也是最关键的一步。其目的是填补缺失的数据、平滑噪声数据、删除冗余数据、纠正错误数据、清除异常数据，将原始的数据格式进行标准化。

2. 数据集成

数据集成是将多个数据源中的数据结合起来并统一存储，建立数据仓库，以更好地解决数据的分布性和异构性问题。数据集成技术的关键是数据高速缓存器。拥有一个包含目标计划、源—目标映射、数据获取、分级抽取、错误恢复和安全性转换的数据高速缓存器，可以大大减少直接访问后端系统和进行复杂实时集成的需求。

3. 数据变换

数据变换是采用线性或非线性的数学变换方法将多维数据压缩成较少维数的数据，消除它们在时间、空间、属性、精度等特征表现方面的差异。数据变换可用相当少的变量捕获原始数据的最大变化，具体变换方法的选择可根据实际数据的属性特点而定，常见的数据变换方法有数据平滑、数据聚焦、数据规范化等。

4. 数据归约

数据归约是指在对数据挖掘任务和数据本身内容理解的基础上寻找依赖于发现目标的数据的有用特征，以缩减数据规模，从而在尽可能保持数据原貌的前提下，最大限度地精简数据量。数据归约主要有两个途径：属性选择和数据采样，分别针对原始数据集中的属性和记录。数据归约技术可以用来得到数据集的归约表示，它虽然小，但仍然大致保持原始数据的完整性。这样，在归约后的数据集上挖掘将更有效，并产生相同(或几乎相同)的

分析结果。数据归约的类型主要有特征归约、样本归约和特征值归约。

2.1.3 数据存储

大数据种类繁多，数据结构化程度不同，传统的结构化数据库无法适应大数据的存储要求。下面介绍 3 种典型的大数据存储方案：分布式文件系统、分布式数据库和云存储。

1. 分布式文件系统

分布式文件系统是指文件系统管理的物理存储资源不一定直接连接在本地节点上，而是通过计算机网络与节点相连，众多的节点组成一个文件系统网络；每个节点可以分布在不同的地点，通过网络进行节点间的通信和数据传输。常见的分布式文件系统有 GFS、HDFS、Lustre、Ceph 等，它们各自适用于不同的领域，其中 GFS 和 HDFS 最具有代表性。GFS 是 Google 公司设计的专用文件系统，主要用于存储海量搜索数据，处理大文件。HDFS 是 Hadoop 分布式文件系统，它是一种被设计成适合运行在通用硬件上的分布式文件系统，具有高容错性的特点。

2. 分布式数据库

分布式数据库是利用网络将物理上分布的多个数据存储单元连接起来组成的逻辑数据库，其基本思想是将集中式数据库中的数据，分散存储到多个数据存储节点上，并通过网络节点连接起来，以获取更大的存储容量和更高的并发访问量。与传统的集中式数据库相比较，分布式数据库具有高扩展性、高并发性、高可用性以及更高的数据访问速度。近年来，随着数据量的高速增长，传统的关系型数据库开始从集中式模型向分布式架构发展，从集中式存储走向分布式存储，从集中式计算走向分布式计算。

3. 云存储

云存储是一种以数据存储和管理为核心的云计算系统，它是指利用集群应用、分布式文件和网络技术系统等功能，通过应用软件协同网络中大量的各种不同类型的存储设备，共同建设一个具有数据存储和业务访问功能的系统，以保证数据的安全性，节约存储空间。互联网技术的发展是实现云存储的基本条件。通过互联网技术，云存储才能实现数据、文档、图片、音频、视频等内容的存储和共享。云存储系统结构主要由存储层、基础管理层、应用接口层、访问层 4 个部分构成。

2.1.4 数据挖掘

数据挖掘是指根据业务的需求和目的，运用合适的工具软件和数据挖掘方法对数据仓库中的数据信息进行处理，寻找出特定的数据规律或数据模式，得出有价值的信息和知识。根据信息存储格式，可以把数据挖掘的对象分为关系数据库、面向对象数据库、数据仓库、文本数据源、多媒体数据库、空间数据库、时态数据库、异质数据库以及 Internet 等。数据挖掘常用的工具软件有：Intelligent Miner、SPSS、SAS、WEKA、Matlab、R 语言、Python 等。数据挖掘的任务是从数据中发现模式，按照数据挖掘的实际作用数据挖掘

任务可分为关联分析、聚类分析、分类、回归、预测、序列和偏差分析。

2.1.5 数据解释

数据解释是一个面向用户的过程，它是指将大数据挖掘及分析结果在显示终端以友好、形象、易于理解的形式呈现给用户。传统的数据解释方法是以文本形式输出结果或者直接在电脑终端上显示结果。大数据分析的结果一般是数据量巨大且关系复杂的结果，传统的分析结果展示方法已基本不可行。现阶段，主要是利用可视化技术、人机交互、数据起源等新的方法将结果展示给用户，帮助用户更加清晰地了解数据处理后的结果，为用户提供决策信息的支持。目前，大部分企业已经引进数据可视化技术和人机交互技术。

1. 数据可视化技术

数据可视化技术主要是通过图形化方法进行清晰、有效的数据传递。其基本思想是使用单个图元元素表示数据库中的每一个数据项，大量的数据集组成数据图像，并以多维数据的形式表示数据的各个属性值。运用可视化技术就可以将数据结果转化为静态或者动态的图形展示给用户，通过交互手段抽取或者集成数据能在画面中动态地显示改变的结果。这样，用户就可以从不同的维度观察数据，对数据进行更深入的观察和分析。可视化技术可以分为 5 类，包括几何技术、图标技术、图形技术、分层技术、混合技术。基于不同的需求可以采取不同的可视化技术，也可以通过多种技术手段来展示数据处理结果。例如，电力网络中电力的传输，为直观地反映各个城市的电力需求状况，可以利用基于图标技术，用不同的颜色标明图中各个城市的电力负载情况。

2. 人机交互技术

人机交互技术是指通过系统输入、输出设备，以有效的方式实现人与系统之间信息交换的技术。其中，系统可以是各类机器、计算机和软件。用户界面或人机界面是人机交互所依托的介质和对话接口，通常包括硬件和软件系统。人机交互技术是一种双向的信息传递过程，既可以由用户向系统输入信息，也可以由系统向用户反馈信息。通过人机交互技术，用户只需要通过输入设备给系统输入有关信息、提示、请示等，系统就会输出或通过显示设备提供相关信息、回答问题等。人机交互技术能够使得大数据分析的数据结果更好地被解释给用户。这种交互式的数据分析过程可以引导用户逐步地进行分析，使得用户在得到结果的同时能够更好地理解分析结果的由来。与此类似的还有数据起源技术，通过该技术可以帮助用户追溯整个数据分析的过程，从而有助于用户理解结果。

@ 2.2 数据来源

要做大数据，首先要了解自己的企业，或者自己所在的行业的核心是什么。也就是说最关键的企业需要找到自己的核心数据(价值)。只有在这个基础上，建立自己的大数据才能做一些延伸。其次，要找到内部的一些外围相关数据，去慢慢地成长它。第一层是核心；第二层是外围相关的数据；第三层是外部机构的一些结构化数据；第四层是社会化

的，以及各种现在所谓的非结构化的数据。第一步，找到核心数据，核心数据现在对很多企业来说实际上就是 CRM，自己的用户系统，这是最重要的。第二步，找到外围数据，通过营销活动等获取大量数据。第三步，找到常规渠道的数据，这就需要企业去找常规渠道里面的数据，跟自己的 CRM 结合起来，才能为下一步做市场营销、做推广、产品创新等建立基础。第四步，找到外部的社会化的或者非结构化的数据，即现在所谓的社会化媒体数据。这方面信息的主要特征是非结构化，而且非常庞大。

下面以金融企业为例，重点讨论金融企业的数据来源、数据现状，企业存在哪些问题以及应该怎么应对。

2.2.1 核心数据

1. 现状

金融企业的核心数据主要有以下几个来源，如图 2.2 所示。

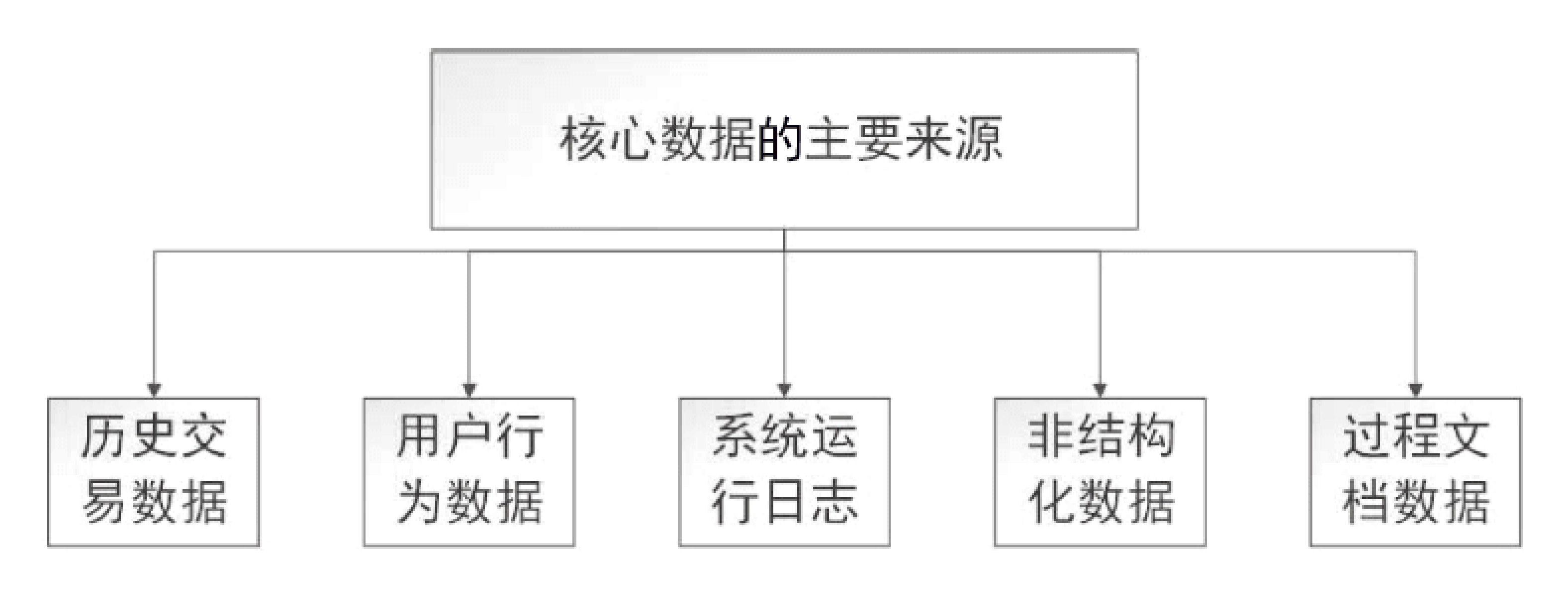


图 2.2　数据来源

1)　历史交易数据

按照主数据的普遍规划来划分，金融企业一般拥有客户数据、交易数据、账户数据等，这些数据有一些已沉淀了多年，伴随着当年的一些金融产品进入数据库，正处于生命周期的某一阶段。这些数据极具潜力价值，通常可以用来促进精确营销、优化产品设计等分析项目。

2)　用户行为数据

企业每天处理海量的交易，有相当一部分交易是网络上的终端客户直接发起的，特别是在一些业务促销活动过程中。因此，柜员服务系统、网上服务系统中产生了大量的业务行为轨迹，这些数据通常可以用来分析提高运营效率、促进精准营销。

3)　系统运行日志

金融企业的应用系统数量较多，分别负责完成各个子领域的业务处理与管理决策。这些应用系统会产生大量的数据库日志和应用程序日志。在日常维护中，这些日志的数量很大、价值密度低，并不受重视。实际上，通过日志分析应用系统效率，是提高应用系统服务水平和客户满意度的重要方法。

4) 非结构化数据

金融企业普遍经济实力雄厚，在众多基础设施建设中投入了巨资。因此，通过大规模的语音呼叫中心、邮件中心、短信中心等客户接触渠道，金融企业拥有发布和采集数据的主动权。另外，不少金融单位有着遍布全国各行政辖区的客户服务大厅，在这里安装了先进的视频监控系统，视频数据既能起到安全防范作用，也能用于分析客户时长等服务类指标。

5) 过程文档数据

金融企业通常都成立了大规模的研发中心和数据中心，按照标准的流程开发和部署应用系统。在这个过程中，将产生大量的需求分析、设计文档、测试报告、上线部署、问题记录等过程和技术文档。这些文档是分析和提升服务水平的重要来源。

2. 问题

核心数据最大的问题在于来源多样、流动性差、共享性差。

1) 数据质量问题

由于某些应用系统开发历史较久，随着架构规划和科学技术的不断进步，导致接口数量多、数据不一致、数据质量差等问题，因此难以进行大数据分析。

2) 内部管理壁垒

金融行业在开展大数据项目获取数据时，最严重的问题是内部管理的壁垒。对于许多企业来说，信息流被各部门彼此分割，数据难以互通，在这种情况下，大数据的共享和汇集变得非常困难，更难以实现大数据的深度应用。

3. 解决方法

数据作为一项资产，部门之间数据壁垒的问题，根源不是各部门造成的，而是公司在数据职责权利的定位方面出现了偏差。

因此，解决此问题需要以下几个途径。

(1) 明确数据相关的职责与归属。金融企业要明确：各个渠道和部门拥有的是数据采集职责，为公司增加数据资产；数据资产的所有权与使用权，只能归公司所有。

(2) 提升对数据资产质量的认识。数据资产至关重要，不少金融企业依靠销售渠道或者第三方平台开展销售，若客户资料质量很差或者根本无法获取，就相当于向公司提供了伪劣的数据资产。

(3) 打通数据流转。金融行业有独立的研发中心和数据中心。其中，研发中心负责程序的开发，不得接触生产数据以及未脱敏的测试数据；数据中心负责程序的部署，不得接触程序源码。应用系统研发与生产的剥离也可能会加剧大数据实施的难度。在大数据这项需要创新与试错的任务面前，数据中心作为数据的实际保有者，往往不愿意向具有创新能力的研发中心提供数据。因此，对大数据应用来说，要确定真正具有创新实践能力的组织架构，并从决策管理层明确所需的各类支持必须到位，确立一定的考核与激励措施，做到利益均沾、成果共享。

2.2.2 外围数据

1. 外围数据的基本准则

(1) 符合法律规定，遵循道德规范。这是一项基本要求。

(2) 在使用外围数据前，分析清楚提供者的商业模式，如果提供者的商业模式会给本企业的未来带来竞争关系，那么合作时需要仔细商榷。

(3) 要在购买与交换之间权衡利弊。在数据所有权不清晰的情况下，交换数据是一种合作举措，可以看作是两家单位以客户为中心的目标下开展的联合行为。

(4) 外部数据的目的是补充内部数据，转化为企业数据资产。如果企业已存在类似的内部数据，但因部门利益割裂的原因无法作为数据资产共享，而采用外购形式弥补，那么这些外部数据往往会变成一个新的分割独占的数据，同样不能变成企业级资产。

2. 外围数据来源

随着数据资产地位的逐渐确立，和固定资产、知识产权一样，围绕着数据的交易会形成新的产业链条。不过数据资产极为特殊，它的价值会随着交换与使用而扩大，这与固定资产、货币资产存在着显著不同。另外，所有权和使用权难以界定，也大大增加了数据交易的难度与风险。

金融企业外围数据的来源如下。

1) 数据共享联盟

对大数据来说，整合和共享的价值更大。例如在医疗行业中，每一个医院对于自己的数据进行分析，需要共享跨医院、跨地域的医疗信息。未来数据将呈现出共享的趋势，数据联盟成为数据集散地之一。

2) 互联网数据

网络爬虫仍然是外部数据的有效获取途径，因为互联网有着最大的数据库。在进行舆情监控时，这类数据来源是不可少的。另外也可以直接和大型互联网平台进行数据交易。

3) 运营商数据

例如，在统计房屋空置率时，利用大数据，根据电力局的智能电表数据、水利局的水表走数、邮局和快递公司的针对该地址的投递率、通信公司的固定电话使用率，基本能找出哪些房屋无人居住。因此，金融企业在寻找优质企业时可以反其道而行之，挖掘客户。未来各行业更好地发展的一条捷径就是客户资源共享。

3. 常见问题

1) 数据获得成本

金融企业数据是非常有价值的一类数据。数据提供商最为知道数据的价值，因此选择通过“购买加交换”的形式提供数据。金融企业需要评估可能付出的成本与代价。

2) 数据价值发挥

很多购买数据的金融企业，是由于内部数据的所有权和使用权不清晰而被迫的行为。在这种情况下，虽然购买数据可以解决某个部门的一时之需，但是这些购入的数据也会陷

入部门壁垒之中，无法最大限度地发挥数据的价值。

2.2.3 常规渠道数据

在大数据时代下，数据将逐步发挥生产资料的作用，数据储备和数据分析能力将成为未来新兴国家最重要的核心战略能力。各地政府正在尝试由信息公开转向数据公开。政府开放数据着重于政府主动开放大量的、实时的、结构化的数据和信息，将其在相关业务上所收集、整理、产生或者保有的数据与信息，主动开放给其他对象(包括社会组织与公众)进行数据创新增值应用。

尽管受格局、意识、管理水平限制，各地各级政府的数据公开呈现出发展迅速但明显不均的态势，但是金融企业应该做好准备，将公开数据资产转化为企业内部的核心竞争能力。

1. 政府数据开放存在内驱动力

在所有数据来源中，政府通常掌握着最大量的、关键性的数据和公共信息资源，如果加大开发力度，将会极大地推动政府办事效率的提升和国家信息服务业的发展。

从政府对内有效管理和对外民生服务两个层面上，降低行政成本、提高决策的科学化水平需要高效、实时的信息系统，而大数据的支持是此类信息系统有效发挥作用的支柱之一。政府提供公共服务、促进经济社会发展的职能发挥同样需要大数据支持。政府掌握了大量关于人口、法人和城市空间地理等数据，如果要提供满足群众需求、有针对性的公共服务，则需要对所掌握的数据进行精细分析。

2. 政府公开数据的步骤

公开数据需要各级政府出台更多具有可操作性的细则和措施。相应部门应制定由政府或者行业协会牵头的整合数据标准。定义政府开放数据的最小数据集，从最小数据集方面来控制收集、扩大开放。然后要制定开放数据的相关法规，界定哪些数据可以开放，因为开放数据有成本，要开放那些最有用、需求量最大的数据。最后，还要加大数据开放所带来的价值分析和评估，研究持续开放的政策。

3. 金融行业积极参与政府数据开放的过程

首先，政府数据公开需要一整套的完整规划、顶层设计和系统建设，贯穿信息收集、整理、存储、发布、服务等全过程，内容包括信息网络、应用系统、信息的采集和发布及相关的管理体制、程序、实施模式和项目管理等。其次，政府公开数据在不同部门、不同层级、不同领域、不同行业之间的分享、交换、整合还存在很多问题，想要建成统一的数据平台，还需要做很多工作。最后，对大数据产业而言，政府公开数据的管理、整合及挖掘，也是具有广阔前景的业务发展方向。

金融行业应秉承社会和政治责任，发挥资金、网点、技术优势，积极参与到政府的数据开放的过程中，以政府为导向，帮助建立起公共数据服务平台，将能够为自身和行业的健康有序发展起到非常重要的基础作用。

2.3 大数据架构

基于上述大数据的特征，通过传统 IT 技术存储和处理大数据成本高昂。一个企业要大力发展大数据应用首先需要解决两个问题：一是低成本、快速地对海量、多类别的数据进行抽取和存储；二是使用新的技术对数据进行分析和挖掘，为企业创造价值。因此，大数据的存储和处理与云计算技术密不可分，在当前的技术条件下，基于分布式系统的 Hadoop，被认为是最适合处理大数据的技术平台。Hadoop 提供的功能：利用服务器集群，根据用户的自定义业务逻辑，对海量数据进行分布式处理。广义上来说，Hadoop 通常是指一个更广泛的概念——Hadoop 生态圈。Hadoop 生态圈如图 2.3 所示。

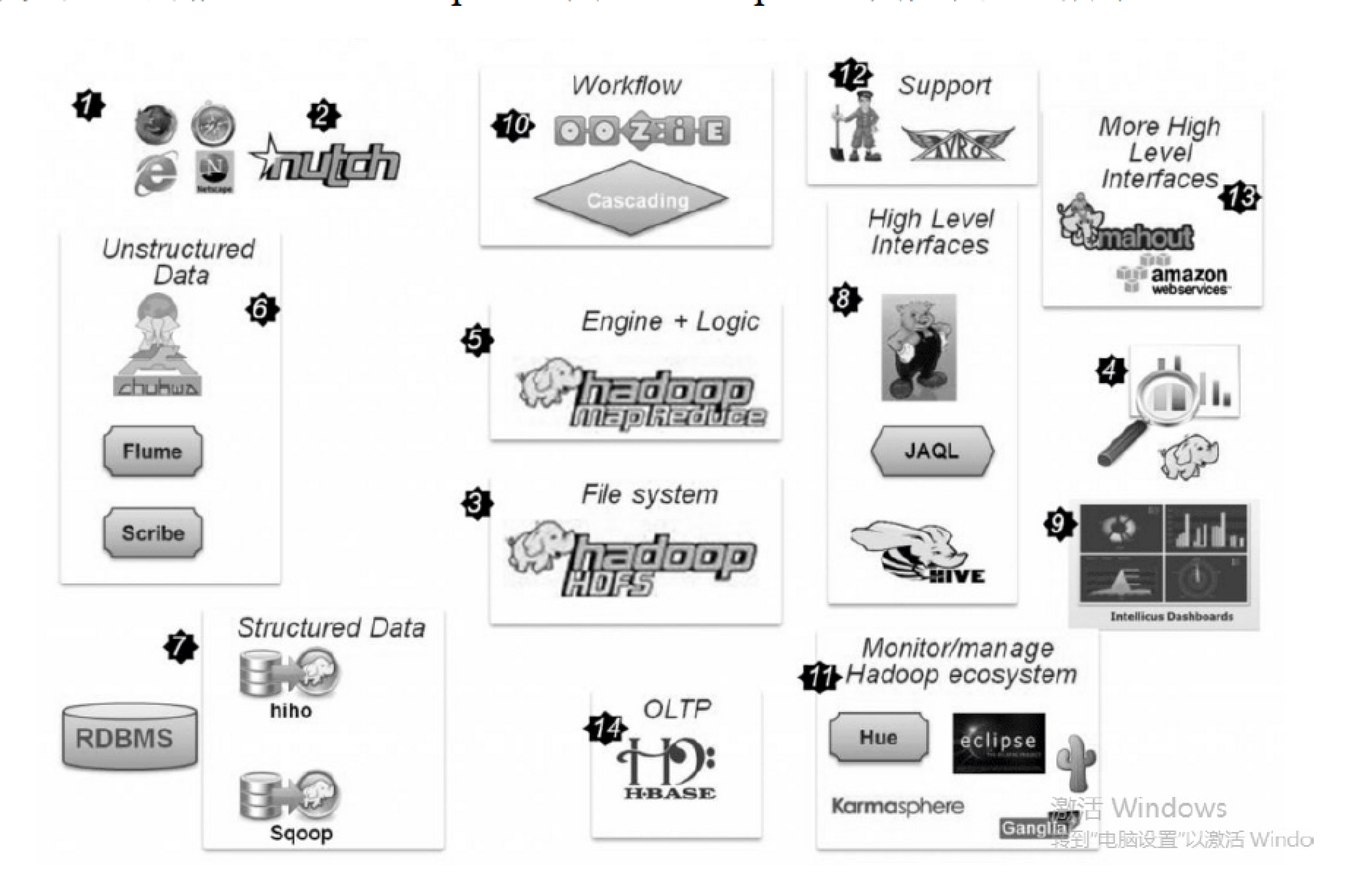


图 2.3 Hadoop 生态圈

各组件简介如下。

1. 主要模块

(1) HDFS：分布式文件系统。

(2) MAPREDUCE：分布式运算程序开发框架，用于大规模数据集的并行计算。

(3) HBASE：基于 Hadoop 的分布式海量数据库，可以将结构化数据文件映射为数据库表，并提供常用的 SQL 支持。Hive 查询引擎将 SQL 语句转化为 Hadoop 平台的 MapReduce 任务运行。

2. 数据管道

(1) Sqoop：主要用于跟关系数据库进行数据交互，通过 JDBC 方式实现数据迁移。

(2) Flume：Cloudera 提供的日志收集框架，用于将海量日志数据并行导入 HDFS 或

者 Hive 中。

(3) DistCp：一般用于在两个 HDFS 集群中传输数据，但目前此命令只支持同版本下集群数据迁移，主要用于冷热数据迁移、测试等场景。

(4) Scribe：Facebook 开源的日志收集系统，它能够从各种日志源上收集日志，存储到一个中央存储系统(可以是 NFS，分布式文件系统等)上，以便于进行集中统计分析处理。它为日志的“分布式收集，统一处理”提供了一个可扩展的、高容错的方案。

3. 数据分析

(1) Hive：提供了一套类数据库的数据存储和处理机制，并采用 HQL (类 SQL)语言对这些数据进行自动化管理和处理。Hive 中的海量结构化数据被看成一个个的表，而实际上这些数据是分布式存储在 HDFS 中的。注意，Hive 是离线查询工具，由于其内部机制，需要把 SQL 转换成 MapReduce 后进行分布式查询，所以最短查询时间也需要十几秒，适用于海量数据场景，不适合即时查询需求。

(2) Impala：Impala 是 Google Dremel 的 Java 实现版本之一。Dermel 由 Google 设计开发，最显著的特性就是支持 SQL 方式在秒级别分析 TB 级别数据(1TB 数据 3 秒完成分析计算)。Impala1.0 版本完全兼容 SQL92 规范，不同于 Hive 将 SQL 转换为 MapReduce 方式，Impala 通过与商用并行关系数据库中类似的分布式查询引框架(由 Query Planner、Query Coordinator 和 Query Exec Engine 三部分组成，与 MR 相似的技术架构，但即时性更好)，可以直接从 HDFS 或者 HBase 中用 SELECT、JOIN 和统计函数查询数据，性能是 Hive(0.81)的 3～90 倍，目前刚发布的 Hive1.0 在原有性能上有很大提升，都属于数据仓库工具，但 Impala 架构更先进。

(3) Pig：Apache Pig 是一个分析大规模数据集的平台，其使用场景和 Hive 相似，Hive 更简单，使用类 SQL 进行数据分析，Pig 使用脚本语言，编程性更强，具体选择主要依靠程序员的熟悉程度及场景复杂度决定。

(4) Mahout：主要用于并行数据挖掘，该框架对目前主流数据挖掘算法都已经基于 MapReduce 进行了实现，节省很多额外开发时间。如推荐引擎、用户关系引擎、GiS 热点聚类都可以基于此框架算法来实现。

(5) Scalding：使用 Scala 编程语言封装 MapReduce 编程模型，支持 DSL(domain-specific language)语法编程，易用性大大提升。主要用于高并发简单 ETL 处理场景。

4. 任务调度

(1) Oozie：其作用就是将多个 MapReduce 作业连接到一起，作为一个工作流程执行。一般情况下，一个大型任务由多个 MapReduce 组成。如果不用 Oozie，需要手动编写大量连接和转换代码，用于串联起多个 MR 任务流程，比较耗时，出错率和维护率也比较高。Oozie 通过 xml 方式配置连接起整个任务流程。与传统工作流引擎作用相似。

(2) Azkaban：美国知名互联公司 Linkedin 发布的开源产品，属于 Oozie 的同类产品，在细节上有区别。

5. 管理

Hue：它是运营和开发 Hadoop 应用的图形化用户界面。对单独的用户来说不需要额外

的安装。另外，Hue 具备简单的权限和用户管理功能，这是其他开源 UI 不具备的。

Hadoop 是一个分布式的基础架构，能够让用户方便高效地利用运算资源和处理海量数据，目前已在很多大型互联网企业得到了广泛应用，如亚马逊、Facebook、Yahoo 等。它是一个开放式的架构，架构成员也在不断扩充完善中。

Hadoop 是一个开发和运行处理大规模数据的软件平台，属于 Apache 开源组织，用 Java 语言开发，用于实现在大量计算机组成的集群中对海量数据进行分布式存储和计算。Hadoop 最核心的设计包含两个模块：HDFS 和 MapReduce。其中 HDFS 提供海量数据的存储，MapReduce 提供海量数据的分布式计算能力。

2.3.1 HDFS 系统

1. HDFS 系统的概念和特性

首先，HDFS 系统是一个文件系统，用于存储文件，通过统一的命名空间——目录树来定位文件。其次，HDFS 系统是分布式的，由很多服务器联合起来实现其功能，集群中的服务器有各自的角色。

HDFS 系统在大数据中的应用是为各类分布式运算框架提供数据存储服务，将大文件、大批量文件，分布式存放在大量的服务器上，以便于采取分而治之的方式对海量数据进行运算分析。

HDFS 系统的特性如下。

(1) 有高容错性的特点。

(2) 整个系统部署在低廉的硬件上。

(3) 提供高传输率来访问应用程序的数据。

(4) 适合超大数据集的应用程序。

(5) 流式数据访问。

HDFS 本身是软件系统，不同于传统硬盘和共享存储介质，在文件操作上有其不同之处。

(1) 不支持文件随机写入。支持随机读，但没有随机写入机制，这与 HDFS 文件写入机制有关，所以不支持断点续传等功能。

(2) 需要客户端与 HDFS 交互。目前已有开源支持 HDFS mount 到 Linux 服务器上，但性能非常不好。

(3) 适合大文件读取场景。因为其分块冗余存储机制，其存储架构在处理小于其分块文件大小的文件时，会浪费管理节点资源，导致效率低。

(4) 吞吐和并发具备横向扩展能力。单节点系统比传统硬盘效率低很多，但在大量机器集群环境下，其吞吐和并发能力可以线性提升，远远高于单一硬件设备。

(5) 不适合高响应系统。由于 HDFS 是为高数据吞吐量应用而设计的，以高延迟为代价。

2. HDFS 的结构

HDFS 中有 3 个重要角色：NameNode、DataNode 和 Client，如图 2.4 所示。

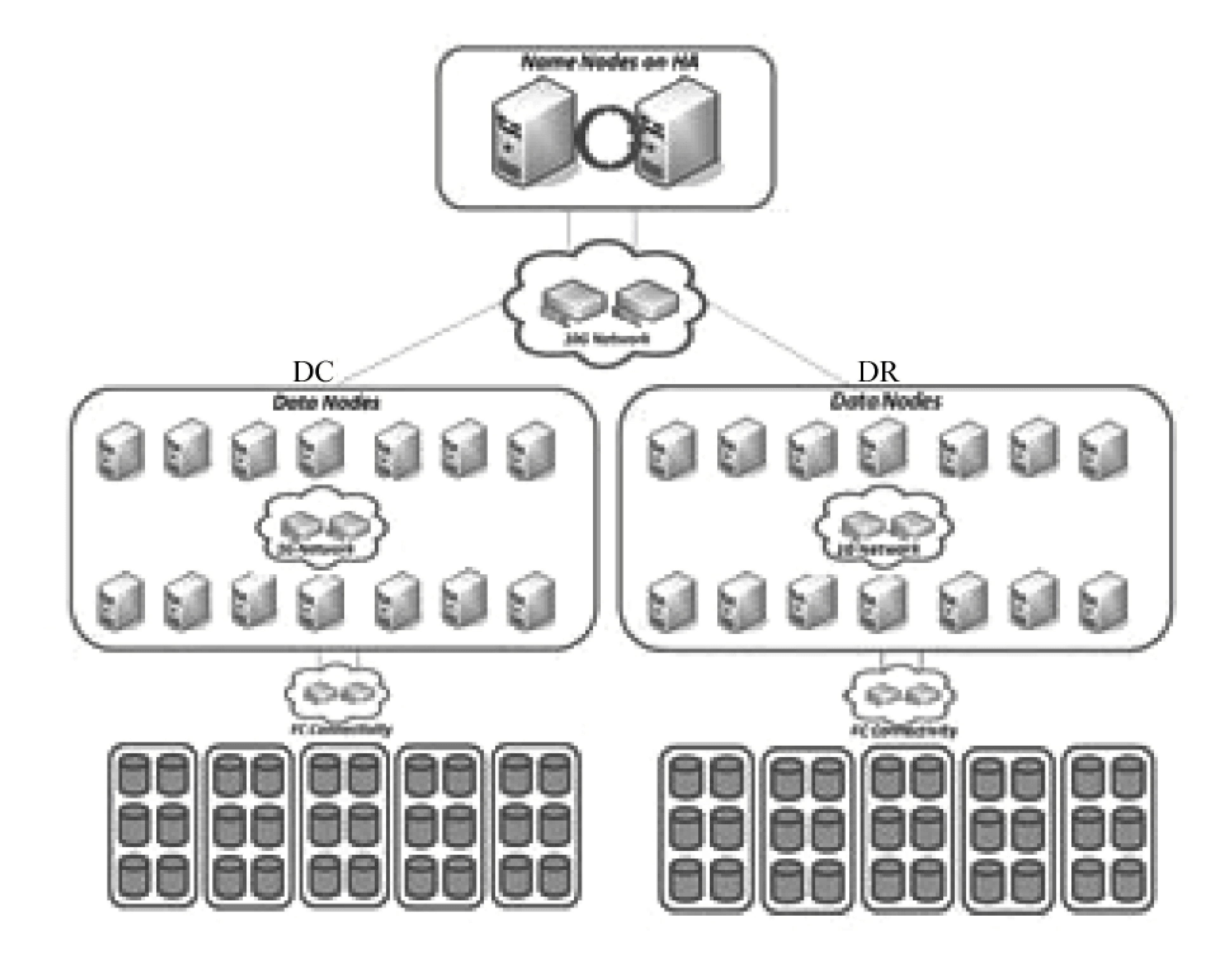


图 2.4 HDFS 结构

对外部客户机而言，HDFS 就像一个传统的分级文件系统，可以删除、移动或重命名文件等。但是 HDFS 架构是基于一组特定的节点构建的，这是由它自身的特点决定的。这些节点包括 NameNode(仅 1 个)，它在 HDFS 内部提供元数据服务；DataNode 为 HDFS 提供存储块。

存储在 HDFS 中的文件被分成块，然后将这些块复制到多台计算机中(DataNode)。这与传统的 RAID 架构大不相同。块的大小(通常为 64MB)和复制的块数量在创建文件时由客户机决定。NameNode 可以控制所有文件操作。HDFS 内部的所有通信都基于标准的 TCP/IP 协议。

1) NameNode

NameNode 是一个通常在 HDFS 实例中的单独机器上运行的软件。它负责管理文件系统名称空间和控制外部客户机的访问。NameNode 决定是否将文件映射到 DataNode 上的复制块上。对于最常见的 3 个复制块，第一个复制块存储在同一机架的不同节点上，最后一个复制块存储在不同机架的某个节点上。Metadata 所有的相关服务都是由 NameNode 提供，包括 filename->block (namespace)，以及 block->DataNode 的对应表。其中，前者通过 FsImage 写入本地文件系统中，而后者是通过每次 HDFS 启动时，DataNode 进行 blockreport 后在内存中重构的数据结构。

实际的 I/O 实务并没有经过 NameNode，只有表示 DataNode 和块的文件映射的元数据经过 NameNode。当外部客户机发送请求要求创建文件时，NameNode 会以块标识和该块的第一个副本的 DataNode 的 IP 地址作为响应。这个 NameNode 还会通知其他将要接收该

块的副本的 DataNode。

NameNode 在一个称为 FsImage 的文件中存储所有关于文件系统名称空间的信息。这个文件和一个包含所有事务的记录文件(EditLog)将存储在 NameNode 的本地文件系统上。FsImage 和 EditLog 文件也需要复制副本，以防文件损坏或 NameNode 系统走失。

2) DataNode

DataNode 也是一个通常在 HDFS 实例中的单独机器上运行的软件。Hadoop 集群中包含一个 NameNode 和大量 DataNode。DataNode 通常以机架的形式组织，机架通过一个交换机将所有系统连接起来。

DataNode 响应来自 HDFS 客户机的读写请求。并且还响应来自 NameNode 的创建、删除和复制块的命令。NameNode 依赖来自每个 DataNode 的定期心跳(Heartbeat)消息。每条消息都包含一个块报告，NameNode 可以根据这个报告验证块映射和其他文件系统元数据。

分布式文件存储的数据节点，存储着文件块(Block)，而文件是由文件块组成的，每个块存储在多个(可配，默认为 3)不同的 DataNode 可以提高数据的可靠性。

如果客户机想将文件写到 HDFS 上，首先需要将文件缓存到本地的临时存储区。如果缓存的数据大于所需的 HDFS 块大小，创建文件的请求将发送给 NameNode。NameNode 将以 DataNode 标识和目标块响应客户机。同时也通知将要保存文件块副本的 DataNode。当客户机开始将临时文件发送给第一个 DataNode 时，将立即通过管道方式将块方式内容转发副本 DataNode。客户机也负责创建保存在相同 HDFS 名称空间中的校验文件。在最后的文件块发送后，NameNode 将文件创建提交到它的持久化数据存储(EditLog 和 FsImage 文件)。

3) Client

用于实现客户端文件存储的所有操作，包括文件的增删以及查询等。

3. HDFS 文件写入与读取

HDFS 文件的写入流程如图 2.5 所示。

(1) 客户端通过 Distributed FileSystem 上的 create()方法指明一个欲创建的文件的文件名，然后 client 通过 RPC 方式与 NameNode 通信创建一个新文件映射关系。

(2) 客户端写数据：FSData OutputStream 把写入的数据分成包(packet)，放入一个中间队列——数据队列(data queue)中去。OutputStream 从数据队列中取数据，同时向 NameNode 申请一个新的 block 来存放它已经取得的数据。NameNode 选择一系列合适的 DataNode(个数由文件的 replication 数决定，默认为 3，构成一个管道线(pipeline)，所以管道线中就有 3 个 DataNode。OutputStream 把数据流式地写入到管道线中的第一个 DataNode 中，第一个 DataNode 再把接收到的数据转到第二个 DataNode 中，以此类推。

(3) FSData OutputStream 同时也维护着另一个中间队列——确认队列(ack queue)，确认队列中的包只有在得到管道线中所有的 DataNode 的确认以后才会被移出确认队列。

(4) 所有文件写入完成后，关闭文件写入流。

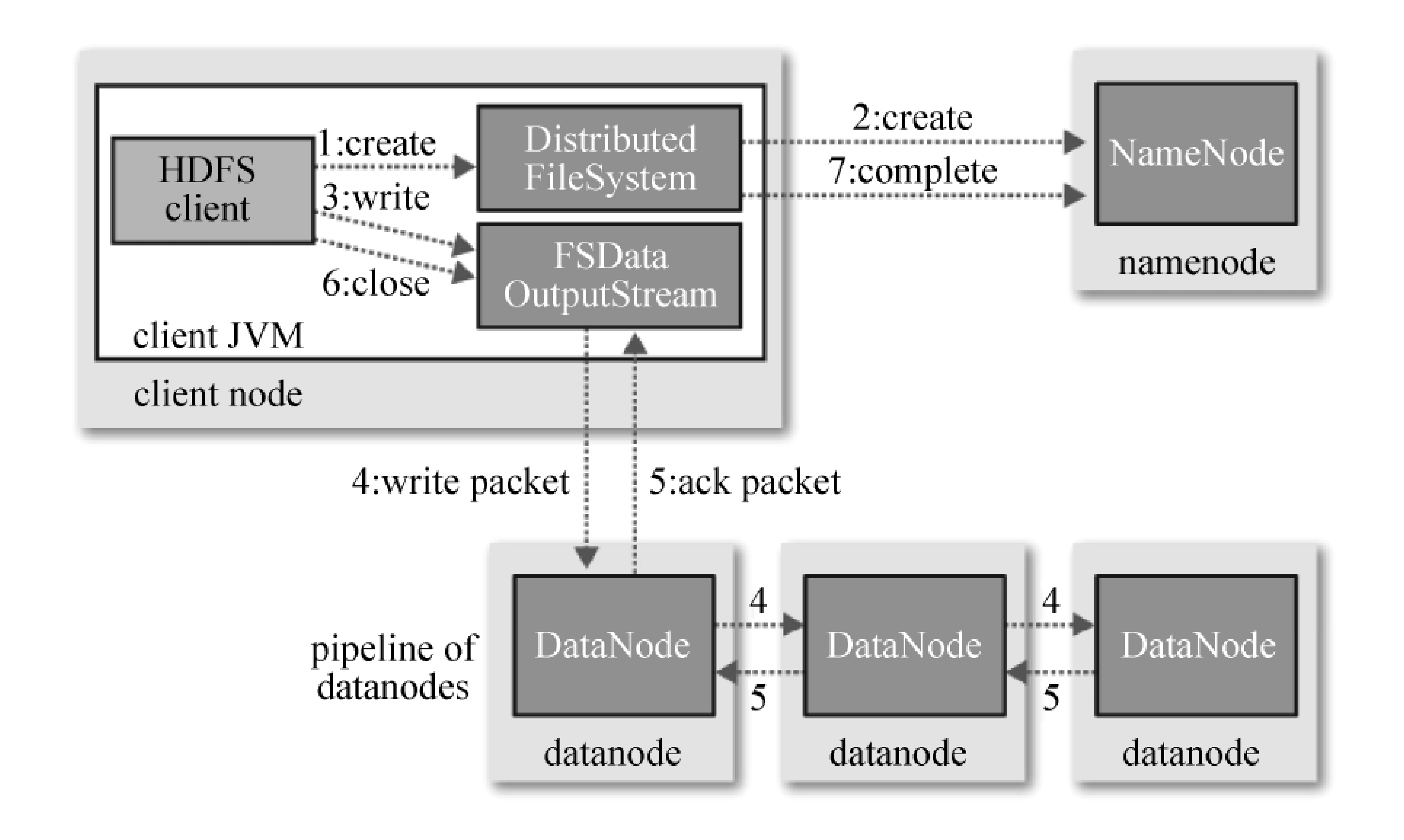


图 2.5　HDFS 文件的写入流程

从以上文件写入流程，可以总结出 HDFS 文件写入具备如下特性。

- 响应时间比较长。
- 文件写入效率与 block 块数和集群数量相关。

HDFS 文件的读取流程如图 2.6 所示。

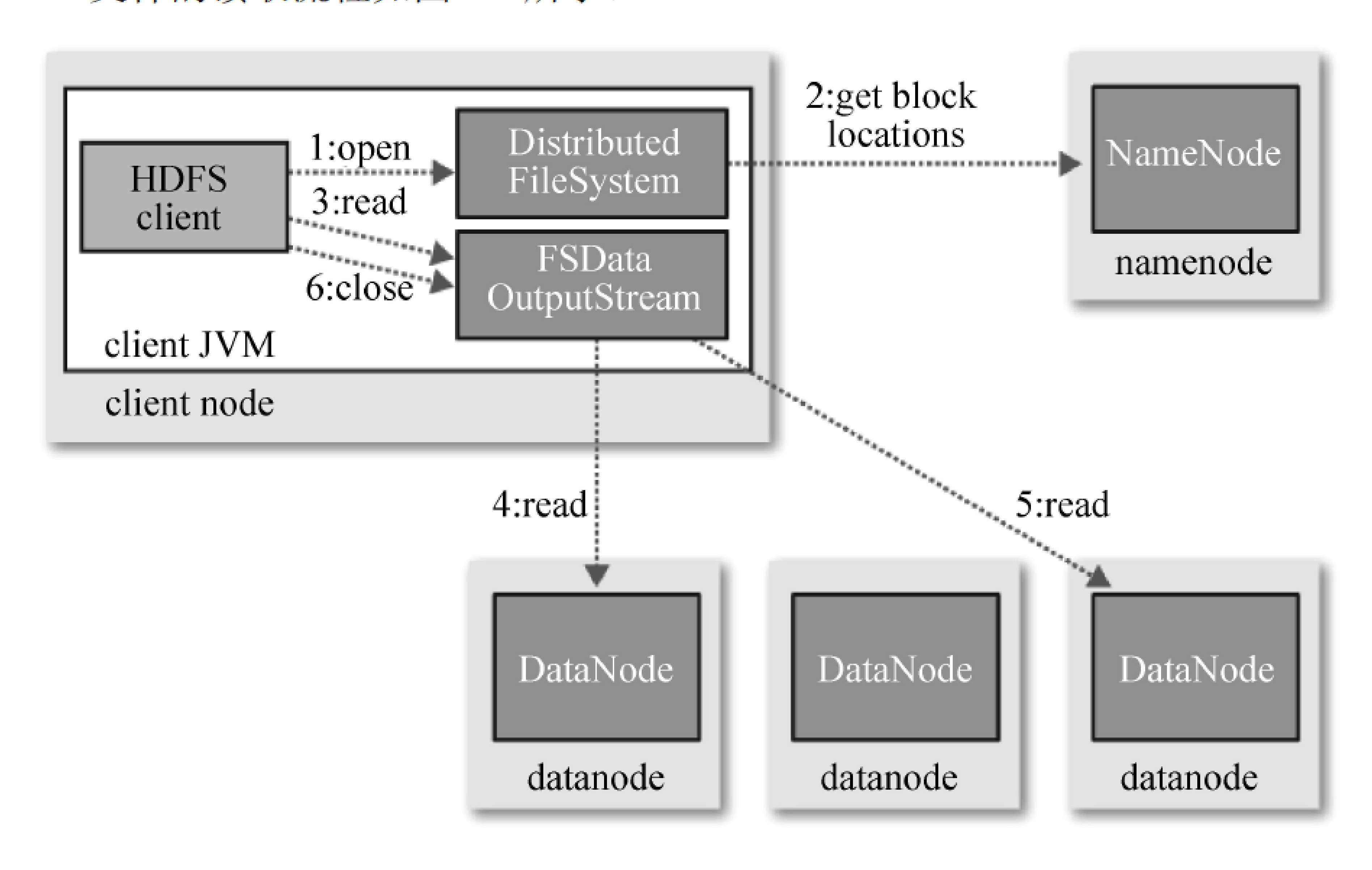


图 2.6　HDFS 文件的读取流程

(1)　打开文件流(open())。

(2)　从 NameNode 读取文件块位置列表。

(3)　FSDataInputSteam 打开 read() 方法。

(4)　根据文件块与 DataNode 的映射关系。

(5) 从不同的 DataNode 中并发读取文件块。

(6) 文件读取完毕，关闭 input 流。

因为冗余机制，当 HDFS 文件读取压力比较大的时候，可以通过提高冗余数的方式，NameNode 可以通过轮询方式，分配不同 client 访问不同 DataNode 上的相同文件块，提升整体吞吐率。

Hadoop 在创建新文件时是如何选择 block 的位置的呢，综合来说，要考虑带宽(包括写带宽和读带宽)和数据安全性，如图 2.7 所示。

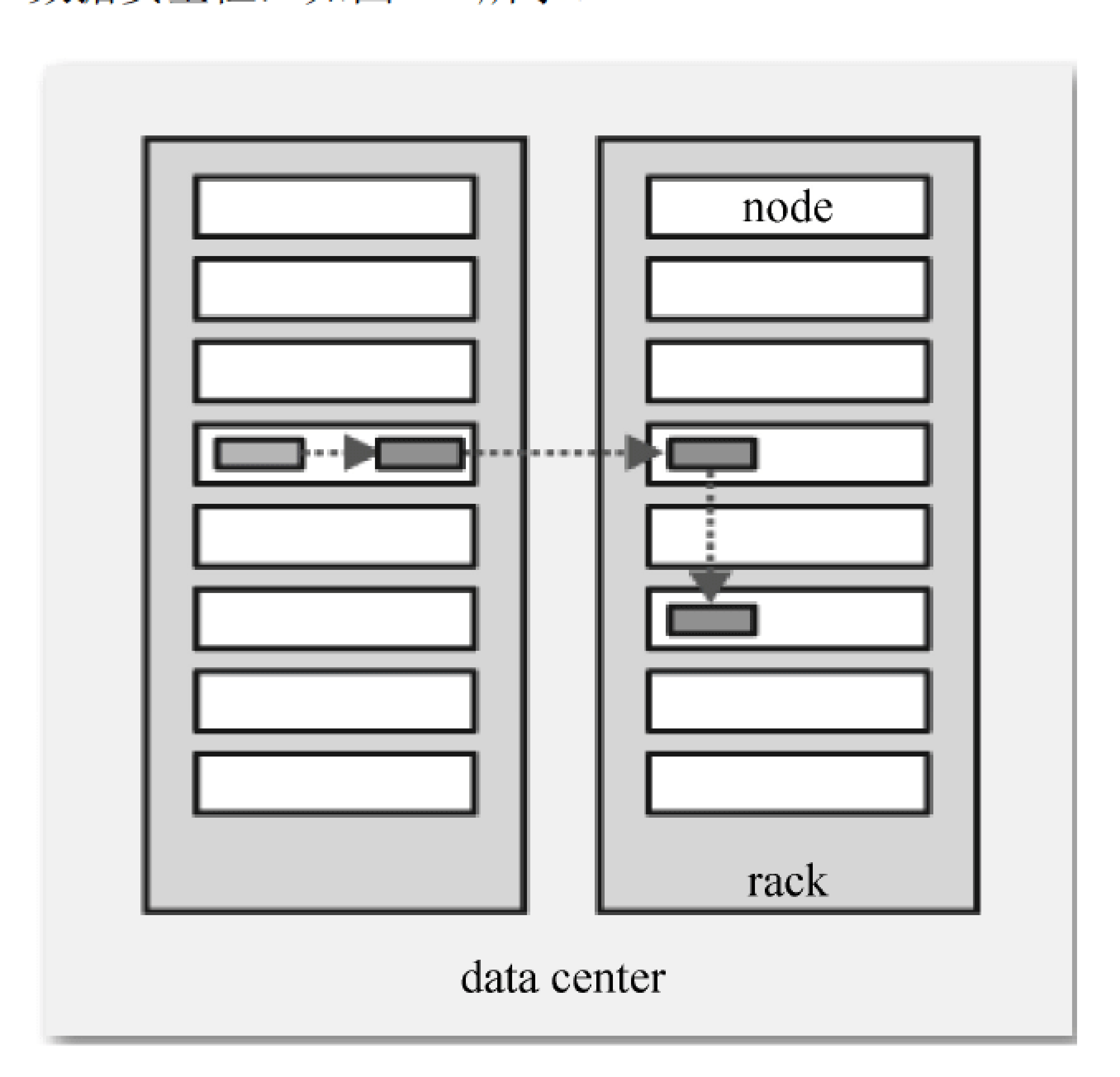


图 2.7 选择 block 的位置

如果把 3 个备份全部放在一个 DataNode 上，虽然可以避免写带宽的消耗，但几乎没有提供数据冗余带来的安全性，如果这个 DataNode 宕机，那么这个文件的所有数据就全部丢失了。另一个极端情况是，如果把 3 个冗余备份全部放在不同的机架上，甚至数据中心里面，虽然这样做数据很安全，但写数据会消耗很多的带宽。HDFS 提供了一个默认备份分配策略：把第一个备份放在与客户端相同的 DataNode 上，第二个放在与第一个不同机架的一个随机 DataNode 上，第三个放在与第二个相同机架的随机 DataNode 上。如果备份数大于 3，则随后的备份在集群中随机存放，Hadoop 会尽量避免过多的备份存放在同一个机架上。

2.3.2 MapReduce

MapReduce 是 Google 提出的并行计算架构，用于大规模数据集(TB 级以上)的并行运算。 此算法的计算能力，随着计算节点的数量增加而线性上升。

图 2.8 表示一个 MapReduce 计算处理思路，可以简要分解为两部分，数据分块映射处理(Map)和数据结果聚合(Reduce)两个步骤，源数据可以存储在 HDFS 或者第三方数据源

上，计算过程临时数据存储在 HDFS 和内存中，最终获得我们需要的计算结果，其具体处理流程如图 2.9 所示。

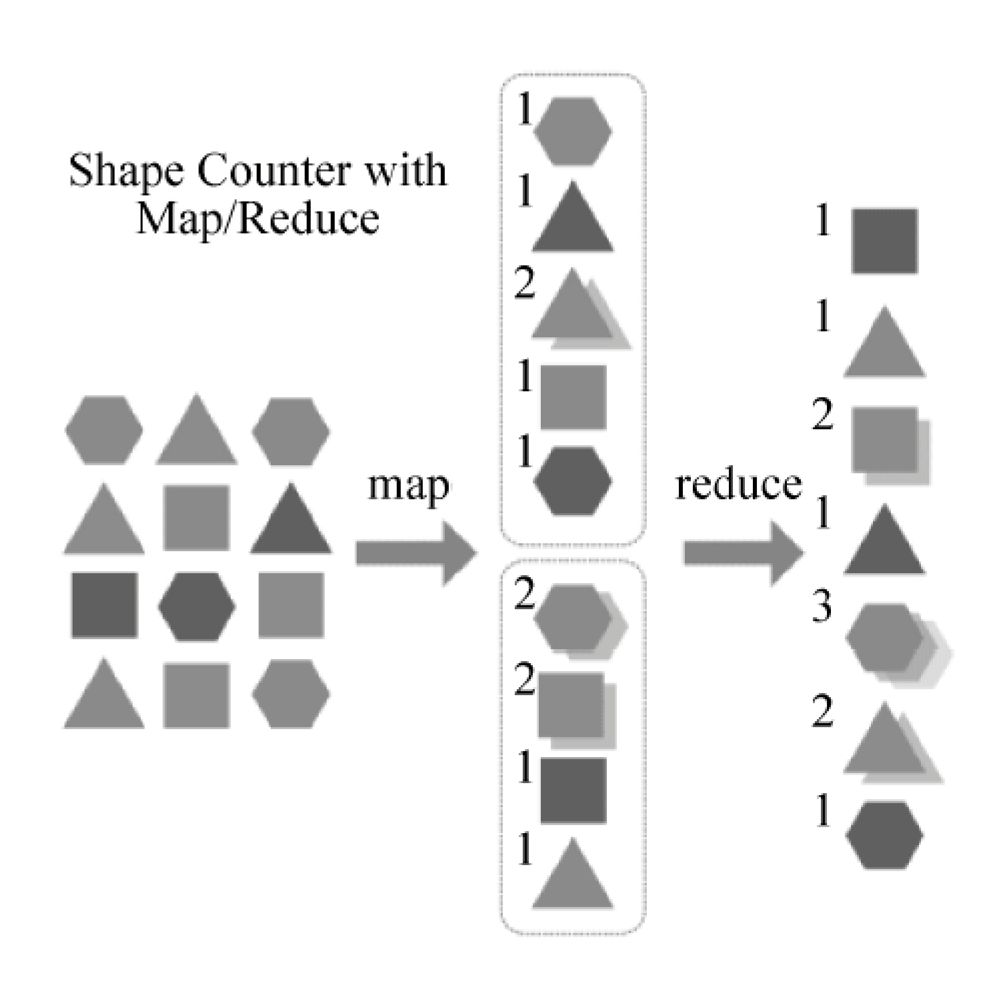


图 2.8 MapReduce 计算处理思路

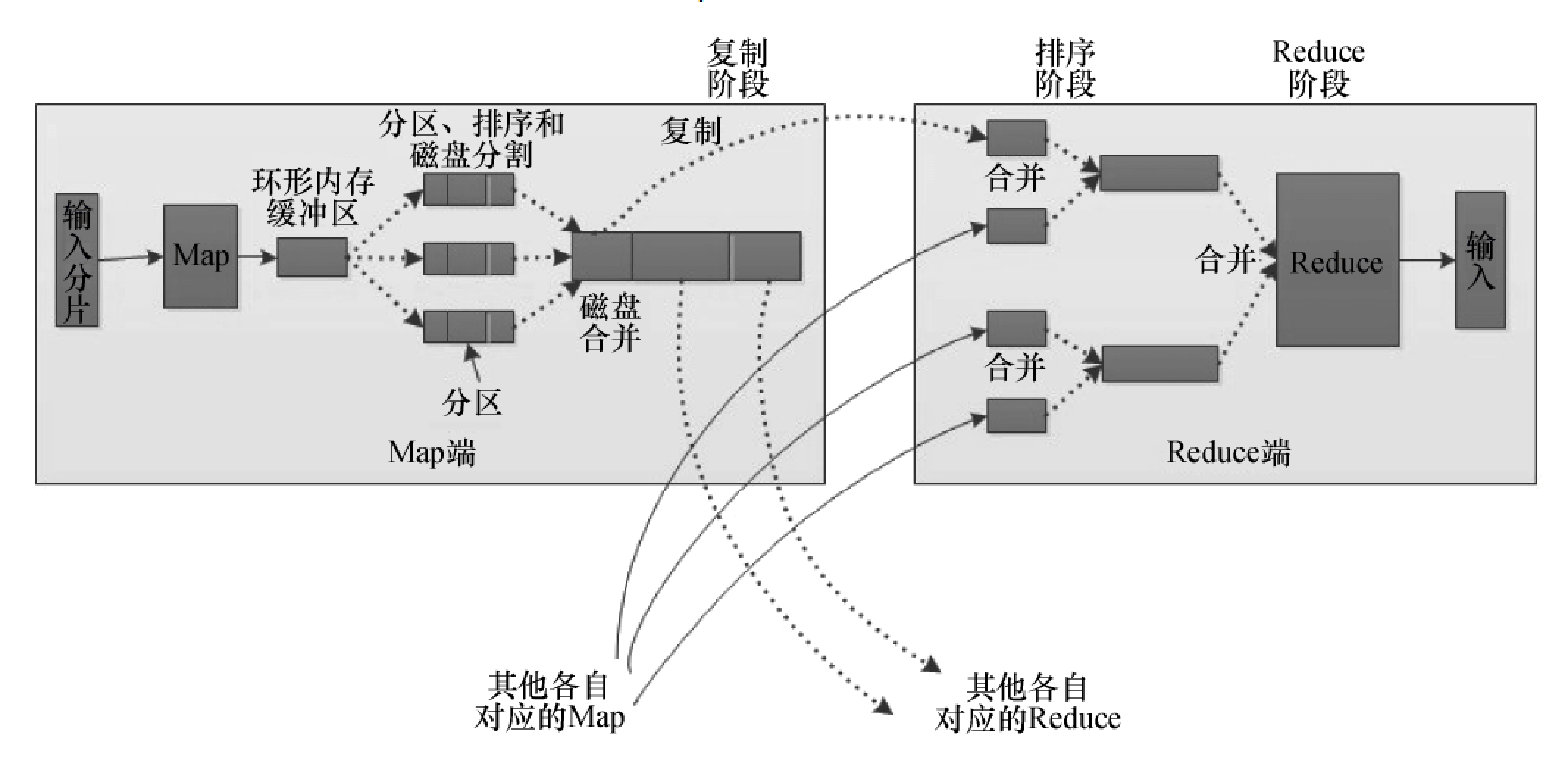


图 2.9 MapReduce 计算处理流程

1. Map 端

(1) 每个输入分片会让一个 Map 任务来处理，默认情况下，以 HDFS 的一个块的大小(默认为 64MB)为一个分片，当然我们也可以自行设置块的大小。Map 输出的结果会暂时放在一个环形内存缓冲区中(该缓冲区的大小默认为 100MB)，当该缓冲区快要溢出时(默认为缓冲区大小的 80%)，会在本地文件系统中创建一个溢出文件，将该缓冲区中的数据写入这个文件。

(2) 在写入磁盘之前，线程首先根据 Reduce 任务的数目将数据划分为相同数目的分区，也就是一个 Reduce 任务对应一个分区的数据。这样做是为了避免有些 Reduce 任务分到大量数据，而有些 Reduce 任务却分到很少数据，甚至没有分到数据的尴尬局面。其实

分区就是对数据进行 Hash 的过程。然后对每个分区中的数据进行排序，如果此时设置了 Combiner，将排序后的结果进行 Map 合并操作，这样做的目的是让尽可能少的数据写入磁盘。

(3) 当 Map 任务输出最后一个记录时，可能会有很多溢出文件，这时需要将这些文件合并。合并的过程中会不断地进行排序和合并操作，目的有两个：①尽量减少每次写入磁盘的数据量；②尽量减少下一复制阶段网络传输的数据量。最后合并成了一个已分区且已排序的文件。为了减少网络传输的数据量，可以将数据压缩。

(4) 将分区中的数据复制给相对应的 Reduce 任务。Map 任务一直和其父 TaskTracker 保持联系，而 TaskTracker 又一直和 JobTracker 保持心跳。所以 JobTracker 中保存了整个集群中的宏观信息。Reduce 任务只需向 JobTracker 获取对应的 Map 输出位置。

2. Reduce 端

(1) Reduce 会接收到不同的 Map 任务传来的数据，并且每个 Map 传来的数据都是有序的。如果 Reduce 端接收的数据量相当小，则直接存储在内存中，如果数据量超过了该缓冲区大小的一定比例，则对数据合并后溢写到磁盘中。

(2) 随着溢写文件的增多，后台线程会将它们合并成一个更大的有序的文件，这样做是为了给后面的合并节省时间。其实不管在 Map 端还是 Reduce 端，MapReduce 都是反复地执行排序、合并操作，这就是为什么有些人会说：排序是 Hadoop 的灵魂。

(3) 合并的过程中会产生许多中间文件(写入磁盘了)，但 MapReduce 会让写入磁盘的数据尽可能地少，并且最后一次合并的结果并没有写入磁盘，而是直接输入到 Reduce 函数。

3. Shuffle

在 Hadoop 的集群环境中，大部分 Map 任务和 Reduce 任务是在不同的 Node 上执行，主要的开销是网络开销和磁盘 I/O 开销，因此 Shuffle 的主要作用如下。

(1) 完整地从 Map 端传输到 Reduce 端。

(2) 跨节点传输数据时，尽可能减少对带宽的消耗(注意是 Reduce 执行的时候去拉取 Map 端的结果)。

(3) 减少磁盘 I/O 开销对任务的影响。

2.3.3 HBase

HBase 是 Google Bigtable 的开源实现版本。数据存储在 HDFS 中，继承了 HDFS 的高可靠性、可伸缩架构，同时自己实现了高性能、列存储、实时读写的特性。

不同于 HDFS 的高吞吐低响应，HBase 设计用于高并发读写场景。

(1) HBase 基于 Hadoop HDFS append 方式进行数据追加操作，非常适合列族文件存储架构。

(2) HBase 写请求，都会先写 redo log，然后更新内存中的缓存。缓存会定期地刷入 HDFS。文件基于列创建，因此任何一个文件(MapFile)只包含一个特定列的数据。

(3) 当某一列的 MapFile 数量超过配置的阈值时，一个后台线程就开始将现有的 MapFile 合并为一个文件，这个操作叫 Compaction。在合并的过程中，读写不会被阻塞。

(4) 读操作会先检查缓存，若未命中，则从最新的 MapFile 开始，依次往最老的 MapFile 找数据。可以想象一次随机读操作可能需要扫描多个文件。

HBase 的文件和日志确实都存储在 HDFS 中，但通过精致设计的算法实现了对高并发数据随机读写的完美支持，这依赖于 HBase 数据排序后存储的特性。与其他的基于 Hash 寻址的 NoSQL 数据库有很大不同。

在使用特性上，原生 HBase 不支持 JDBC 驱动，也不支持 SQL 方式进行数据查询，只有简单的 PUT 和 GET 操作。数据查询通过主键(row key)索引和 Scan 查询方式实现，在事务上，HBase 支持单行事务(可通过上层应用和模块如 hive 或者 coprocessor 来实现多表 join 等复杂操作)。HBase 主要用来存储非结构化和半结构化的松散数据，如图 2.10 所示。

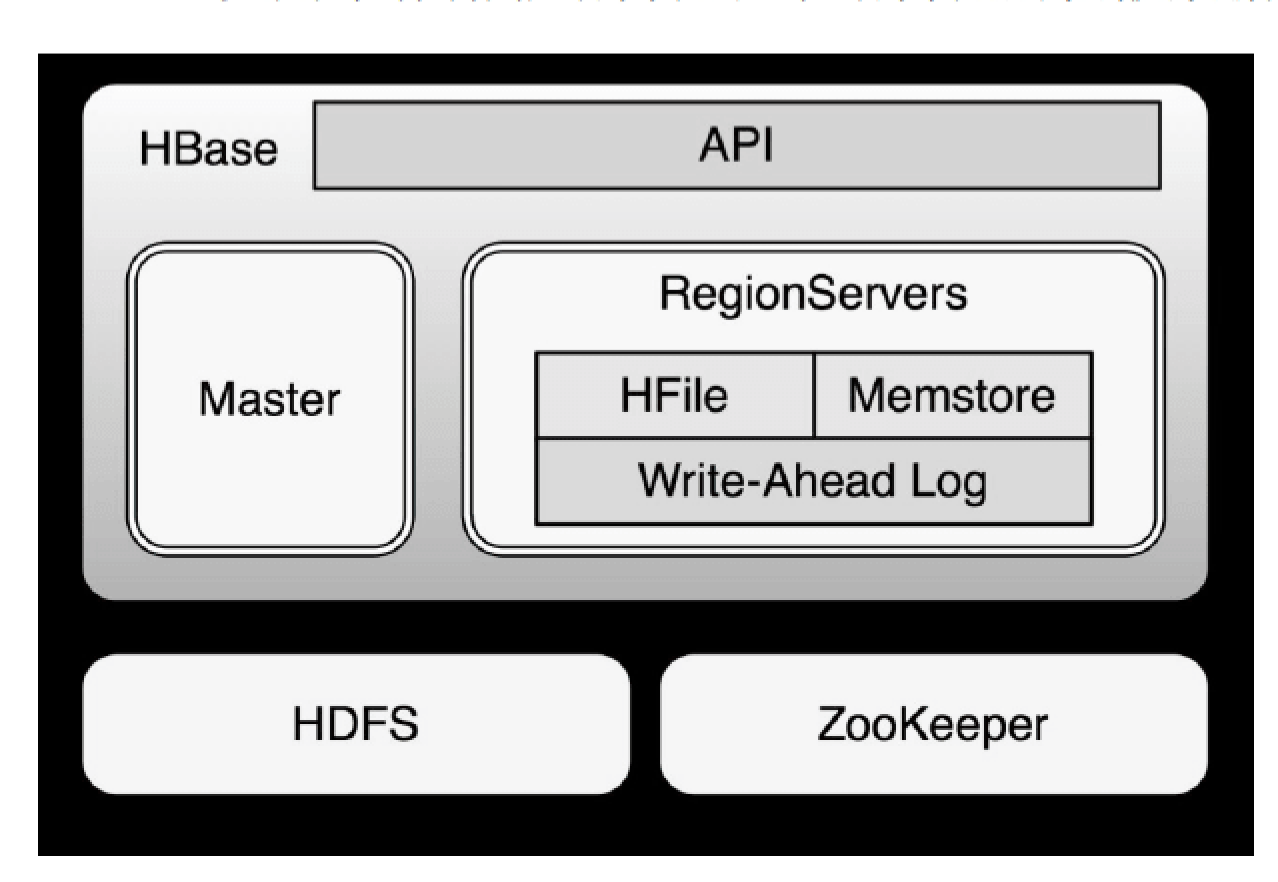


图 2.10 HBase 的架构

HBase 中的表一般有以下特点。

(1) 大：一个表可以有上亿行，上百万列。

(2) 面向列：面向列(族)的存储和权限控制，列(族)独立检索。

(3) 稀疏：对于为空(null)的列，并不占用存储空间(每个列族是一个文件，没内容的情况下不会占用空间)，因此，表可以设计得非常稀疏。

(4) HBase 适用于海量高并发文本数据写入、存储、查询需求场景，这些数据量是传统数据库难以满足的，以下列了一些适用场景。

(5) 详单管理、查询。

(6) GiS 数据存储、统计。

2.4 数据挖掘方法

在大数据时代，数据挖掘是最关键的工作。大数据的挖掘是从海量的、不完全的、有噪声的、模糊的、随机的大型数据库中发现隐含在其中的有价值的、潜在有用的信息和知识的过程，也是一种决策支持过程。其主要基于人工智能、机器学习、模式学习、统计学

等。通过对大数据高度自动化的分析，做出归纳性的推理，从中挖掘出潜在的模式，可以帮助企业、商家、用户调整市场政策、减少风险、理性面对市场，并做出正确的决策。目前，在很多领域尤其是在商业领域(如银行、电信、电商等)，数据挖掘可以解决很多问题，包括市场营销策略制定、背景分析、企业管理危机等。大数据的挖掘常用的方法有分类、回归分析、聚类分析、关联规则、因子分析、主成分分析、神经网络方法、Web 数据挖掘等。这些方法从不同的角度对数据进行挖掘。

2.4.1 分类分析

分类是数据挖掘技术中运用最为广泛也是比较重要的分析手段，它是指运用训练数据集，通过分析数据的特征和运用一定的算法求得分类规则，该分类规则就是数据分类的模型，然后运用该模型对任何位置的数据对象进行分类。分类分为两个阶段：①构建分类模型，通过一定的算法对已知类标记的数据集建立分类模型；②用第一阶段构造的模型来预测给定的数据对象的类别。比较典型的分类方法有决策树分类方法、神经网络分类法、贝叶斯分类法以及 K-近邻分类法。分类分析可以被用于分析客户的属性和特征，进行精准营销。

1. 决策树

决策树是用于分类和预测的主要技术之一，决策树学习是以实例为基础的归纳学习算法，它着眼于从一组无次序、无规则的实例中推理出以决策树表示的分类规则。构造决策树的目的是找出属性和类别间的关系，用它来预测将来未知类别的记录的类别。它采用自顶向下的递归方式，在决策树的内部节点进行属性的比较，并根据不同的属性值判断从该节点向下的分支，在决策树的叶节点得到结论。决策树的表现形式类似于流程图的树结构，在决策树的内部节点进行属性值测试，并根据属性值判断由该节点引出的分支，在决策树的叶节点得到结论。内部节点是属性或者属性组合，而叶节点代表样本所属的类或类分布。经由训练样本集产生一棵决策树后，为了对未知样本集进行分类，需要在决策树上测试未知样本的属性值。测试路径是由根节点到某个叶节点，叶节点代表的类就是该样本所属的类。

2. 贝叶斯分类

贝叶斯(Bayes)分类算法是利用统计学贝叶斯定理，来预测类成员的概率，即给定一个样本，计算该样本属于一个特定的类的属性。这些算法主要利用 Bayes 定理来预测一个未知类别的样本属于各个类别的可能性，选择其中可能性最大的一个类别作为该样本的最终类别。由于贝叶斯定理的成立本身需要一个很强的条件独立性假设前提，而此假设在实际情况中经常是不成立的，因而其分类准确性就会下降。为此就出现了许多降低独立性假设的贝叶斯分类算法，如 TAN 算法，它是在贝叶斯网络结构的基础上增加属性对之间的关联来实现的。

贝叶斯分类的主要算法包括朴素贝叶斯分类算法、贝叶斯网络分类算法等。

朴素贝叶斯分类(Naïve Bayes Analysis，NBC)，假设每个属性之间都是相互独立的，

并且每个属性对非类问题产生的影响都是一样的，即一个属性值对给定类的影响独立于其他属性的值。

贝叶斯定理是概率论中的一个结果，它跟随机变量的条件概率以及边缘概率分布有关。通常来讲，事件 A 在事件 B 发生的条件下的概率，与事件 B 在事件 A 发生的条件下的概率是不一样的，这两者有确定的关系，贝叶斯定理就是这种关系的陈述。

3. k-近邻分类法

k-近邻分类法不是事先通过数据来选好分类模型，再对未知样本分类，而是存储带有标记的样本集，给一个没有标记的样本，用样本集中 k 个与之相近的样本对其进行即时分类。k-近邻就是找出 k 个相似的样本来建立目标函数逼近。

k-近邻的基本思路：首先，存储一些标记好的样本集；其次，要有一个未知类的样本用来对其分类；其次，逐一取出样本集中的样本，与未知类样本相比较，找到 k 个与之相近的样本，用这 k 个样本的多数的类为未知样本定类；最后，在样本集为连续值时，用 k 个样本的平均值为未知样本定值。

2.4.2 回归分析

回归分析是指对具有相关关系的两个变量或多个变量建立合适的数学模型，以近似地表示变量之间平均变化关系的一种统计方法。回归分析与分类分析类似，但回归分析的目的不是寻找描述类的模式，而是寻找变量间的关系模式以确定数值。例如简单的线性回归技术，它的结果是一个函数，可以根据输入变量的值来计算输出变量的值。比较流行的回归分析技术有线性回归和逻辑回归，两者的区别在于线性回归的因变量是连续的，逻辑回归的变量是离散的。此外，还有非线性回归模型，有的可以转化为线性模型。回归分析方法被广泛地用于解释市场占有率、销售额、品牌偏好及市场营销效果。

1. 线性回归

线性回归是利用数理统计中的回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛。

线性回归有很多实际用途，分为以下两大类。

(1) 如果目标是预测或者映射，线性回归可以用来对观测数据集的值和 X 的值拟合出一个预测模型。当完成这样一个模型以后，对于一个新增的 X 值，在没有给定与它相配对的 y 值的情况下，可以用这个拟合过的模型预测出一个 y 值。

(2) 给定一个变量 y 和一些变量 $X_1,\ldots,X_p$，这些变量有可能与 y 相关，线性回归分析可以用来量化 y 与 X_j 之间相关性的强度，评估出与 y 不相关的 X_j，并识别出哪些 X_j 的子集包含关于 y 的冗余信息。

2. Logistic 回归分析

Logistic 回归模型是一种概率模型，适合于病例—对照研究、随访研究和横断面研究，且结果发生的变量取值必须是二分的或多项分类。可用影响结果变量发生的因素作为自变量与因变量，建立回归方程。

Logistic 回归分析的主要用途：一是寻找危险因素；二是预测；三是判别。

2.4.3 其他方法

1. 聚类分析

聚类分析源于许多研究领域，包括数据挖掘、统计学、机器学习、模式识别等。聚类分析是指将物理或抽象对象的集合分组成为由类似的对象组成的多个类的分析过程。聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。聚类分析是一种探索性的分析，在分类的过程中，人们不必事先给出一个分类的标准，聚类分析能够从样本数据出发，自动进行分类。聚类分析所使用方法的不同，常常会得到不同的结论。不同研究者对于同一组数据进行聚类分析，所得到的聚类数未必一致。作为数据挖掘中的一个功能，聚类分析能作为一个独立的工具来获得数据分布的情况，并且概括出每个簇的特点，或者集中注意力对特定的某些簇做进一步分析。数据挖掘技术的一个突出特点是能处理巨大的、复杂的数据集，这对聚类分析技术提出了特殊的挑战，要求算法具有可伸缩性、可处理不同类型的属性、可发现任意形状的类及处理高维数据等。根据潜在的各项应用，数据挖掘对聚类分析方法提出了不同要求。

聚类类似于分类，但与分类的目的不同，是针对数据的相似性和差异性将一组数据分为几个类别。属于同一类别的数据间的相似性很大，但不同类别之间数据的相似性很小，跨类的数据关联性很低。

聚类在数据挖掘中的典型应用有以下 3 个方面。①聚类分析可以作为其他算法的预处理步骤：利用聚类进行数据预处理，可以获得数据的基本情况，在此基础上进行特征抽取或分类可以提高精确度和挖掘效率。也可将聚类结果用于进一步关联分析，以获得进一步的有用信息。②可以作为一个独立的工具来获得数据的分布情况：聚类分析是获得数据分布情况的有效方法。通过观察聚类得到每个簇的特点，可以集中对特定的某些簇做进一步的分析。③聚类分析可以完成孤立点挖掘。许多数据挖掘算法试图使孤立点影响最小化，或者排除它们。然而孤立点本身可能是非常有用的，如在金融欺诈探测中，孤立点可能预示着金融欺诈行为的存在。

聚类分析法有快速聚类和系统聚类。

1) 快速聚类

要求事先确定分类。它不仅要求确定分类的类数，而且还需要事先确定点，也就是聚类种子，然后，根据其他点离这些种子的远近把所有点进行分类。再然后就是将这几类的中心(均值)作为新的基石，再分类。如此迭代。

2) 系统聚类

系统聚类是将样品分成若干类的方法，其基本思想是：先将每个样品各看成一类，然后规定类与类之间的距离，选择距离最小的一对合并成新的一类，计算新类与其他类之间的距离，再将距离最近的两类合并，这样每次减少一类，直至所有的样品合为一类为止。

2. 关联规则

关联规则挖掘是数据挖掘中研究较早而且至今仍活跃的研究方法之一。关联规则是隐藏在数据项之间的关联或相互关系，即可以根据一个数据项的出现推导出其他数据项的出现。关联规则的挖掘过程主要包括两个阶段：第一阶段为从海量原始数据中找出所有的高频项目组；第二阶段为从这些高频项目组产生关联规则。关联规则挖掘技术已经被广泛应用于金融行业企业中用以预测客户的需求，通过捆绑客户可能感兴趣的信息供用户了解并获取相应信息来改善自身的营销。

关联规则是描述数据库中数据项之间所存在的关系的规则，即根据一个事务中某些项的出现可导出另一些项在同一事务中也出现，即隐藏在数据间的关联或相互关系。

在客户关系管理中，通过对企业的客户数据库里的大量数据进行挖掘，可以从大量的记录中发现有趣的关联关系，找出影响市场营销效果的关键因素，为产品定位、定价与定制客户群，客户寻求、细分与保持，市场营销与推销，营销风险评估和诈骗预测等决策支持提供参考依据。

(1) Apriori 算法：使用候选项集找频繁项集。

Apriori 算法是一种最有影响的挖掘布尔关联规则频繁项集的算法。其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则。在这里，所有支持度大于最小支持度的项集称为频繁项集，简称频集。

该算法的基本思想是：首先找出所有的频集，这些项集出现的频繁性至少和预定义的最小支持度一样。由频集产生强关联规则，这些规则必须满足最小支持度和最小可信度。然后使用第 1 步找到的频集产生期望的规则，产生只包含集合的项的所有规则，其中每一条规则的右部只有一项，这里采用的是中规则的定义。一旦这些规则被生成，那么只有那些大于用户给定的最小可信度的规则才被留下来。为了生成所有频集，使用了递推的方法。

可能产生大量的候选集，以及可能需要重复扫描数据库，是 Apriori 算法的两大缺点。

(2) 基于划分的算法。

Savasere 等设计了一个基于划分的算法。这个算法先把数据库从逻辑上分成几个互不相交的块，每次单独考虑一个分块并对它生成所有的频集，然后把产生的频集合并，用来生成所有可能的频集，最后计算这些项集的支持度。这里分块的大小选择要使得每个分块可以被放入主存，每个阶段只需被扫描一次。而算法的正确性是由每一个可能的频集至少在某一个分块中是频集保证的。该算法是可以高度并行的，可以把每一分块分别分配给某一个处理器生成频集。产生频集的每一个循环结束后，处理器之间进行通信来产生全局的候选 k-项集。通常这里的通信过程是算法执行时间的主要瓶颈；而另一方面，每个独立的处理器生成频集的时间也是一个瓶颈。

(3) FP-树频集算法。

针对 Apriori 算法的固有缺陷，J. Han 等提出了不产生候选挖掘频繁项集的方法：FP-树频集算法。采用分而治之的策略，在经过第一遍扫描之后，把数据库中的频集压缩进一

棵频繁模式树(FP-tree)，同时依然保留其中的关联信息，随后再将 FP-tree 分化成一些条件库，每个库和一个长度为 1 的频集相关，然后再对这些条件库分别进行挖掘。当原始数据量很大的时候，也可以结合划分的方法，使得一个 FP-tree 可以放入主存中。

3. 因子分析

因子分析的基本目的就是用少数几个因子描述许多指标或因素之间的联系，即将相关比较密切的几个变量归在同一类中，每一类变量就成为一个因子，以较少的几个因子反映原资料的大部分信息。

运用这种研究技术，我们可以方便地找出影响消费者购买、消费和满意度的主要因素是哪些，以及这些因素的影响力如何。运用这种研究技术，我们还可以为市场细分做前期分析。

4. 主成分分析

设法将原来的变量重新组合成一组新的互相无关的几个综合变量，同时根据实际需要从中可以取出几个较少的综合变量尽可能多地反映原来变量的信息的统计方法叫作主成分分析或称主分量分析，这也是数学上用来降维的一种方法。

主成分分析是设法将原来众多的具有一定相关性(比如 P 个指标)，重新组合成一组新的互相无关的综合指标来代替原来的指标。

最经典的做法就是用 F1(选取的第一个线性组合，即第一个综合指标)的方差来表达，即 Var(F1)越大，表示 F1 包含的信息越多。因此，在所有的线性组合中选取的 F1 应该是方差最大的，故称 F1 为第一主成分。如果第一主成分不足以代表原来 P 个指标的信息，再考虑选取 F2，即选第二个线性组合，为了有效地反映原来的信息，F1 已有的信息就不需要再出现在 F2 中，用数学语言表达就是要求 Cov(F1,F2)=0，则称 F2 为第二主成分，以此类推可以构造出第三、第四……第 P 个主成分。

主成分分析作为基础的数学分析方法，其实际应用十分广泛，比如人口统计学、数量地理学、分子动力学模拟、数学建模、数理分析等学科中均有应用，是一种常用的多变量分析方法。

5. 神经网络方法

神经网络作为一种先进的人工智能技术，因其自身自行处理、分布存储和高度容错等特性非常适合处理非线性的以及那些以模糊、不完整、不严密的知识或数据为特征的处理问题，它的这一特点十分适合解决数据挖掘的问题。典型的神经网络模型主要分为三大类：第一类是用于分类预测和模式识别的前馈式神经网络模型，其主要代表为函数型网络、感知机。第二类是用于联想记忆和优化算法的反馈式神经网络模型，以 Hopfield 的离散模型和连续模型为代表。第三类是用于聚类的自组织映射方法，以 ART 模型为代表。虽然神经网络有多种模型及算法，但在特定领域的数据挖掘中使用何种模型及算法并没有统一的规则，而且人们很难理解网络的学习及决策过程。

6. Web 数据挖掘

Web 数据挖掘是一项综合性技术，指 Web 从文档结构和使用的集合 C 中发现隐含的

模式 P，如果将 C 看作是输入、P 看作是输出，那么 Web 挖掘过程就可以看作是从输入到输出的一个映射过程。

当前越来越多的 Web 数据都是以数据流的形式出现的，因此对 Web 数据流挖掘就具有很重要的意义。目前常用的 Web 数据挖掘算法有：PageRank 算法，HITS 算法以及 LOGSOM 算法。这 3 种算法提到的用户都是笼统的用户，并没有区分用户的个体。目前 Web 数据挖掘面临着一些问题，主要包括：用户的分类问题，网站内容时效性问题，用户在页面停留时间问题，页面的链入与链出数问题等。在 Web 技术高速发展的今天，这些问题仍旧值得研究并加以解决。

7. 序列分析

序列分析是对序列数据进行分析以发现蕴藏其中的模式和规律。序列数据和时间序列数据都是连续的观测值，观测值之间相互依赖。它们之间的差别在于序列数据包含离散的状态，而时间序列是连续的数值。序列数据和关联数据比较相似，它们都是一个项集或一组状态，区别在于序列分析分析的是状态的转移，将数据间的关联性和时间联系起来，而关联分析不需要考虑时间问题。Markov 链是进行序列分析的主要技术之一。

8. 偏差分析

数据库中一般存在着很多异常数据，找出这些异常数据非常重要，偏差分析可以解决此类问题。偏差分析用于检测数据现状、历史记录与标准之间的显著变化和偏离，例如，观测结果与期望的偏离、分类中的反常实例、模式的例外等。偏差分析的基本方法就是寻找观察结果与参照之间的差别。例如，信用卡欺诈案行为检测、网络入侵检测、劣质产品分析等。

9. 预测

预测是大数据最核心的功能。大数据预测是指运用历史数据和预测模型预测未来某件事情的概率。精度和不确定性是预测的关注点，通常用预测方差进行衡量。预测技术是以表示一系列时间值的数列作为输入，接下来运用计算机学习和统计技术对数据进行周期性分析、趋势分析和噪声分析，进而估算这些序列未来的值。例如，可以通过挖掘企业的历史销售数据预测该企业未来一年的销售额。

本章总结

- 大数据的处理流程归纳为：首先是利用多种轻型数据库收集海量数据，对不同来源的数据进行预处理后，整合存储到大型数据库中，然后根据企业或个人目的和需求，运用合适的数据挖掘技术提取有益的知识，最后利用恰当的方式将结果展现给终端用户。具体包括：数据采集、数据预处理、数据存储、数据挖掘以及数据解释这五个步骤。
- 要做大数据，首先要了解自己的企业。第一步，找到核心数据。第二步，获取外围数据，通过营销活动等获取大量数据。第三步，常规渠道的数据，这就需要企

业去找常规渠道里面的数据，跟自己的 CRM 结合起来。第四步，获取外部的社会化的或者非结构化的数据，即现在所谓的社会化媒体数据。这方面信息的主要特征是非结构化，而且数量庞大。

- 金融企业的核心数据主要来源：历史交易数据、用户行为数据、系统运行日志、非结构化数据、过程文档数据。核心数据最大的问题在于来源多样、流动性差、共享性差。要解决这些问题，必须要明确数据相关的职责与归属、提升对数据资产质量的认识和打通数据流转。
- 金融企业外围数据主要来源：数据共享联盟、互联网数据、运营商数据。外围数据存在的问题主要是：存在数据获得成本以及无法最大限度地发挥数据的价值。
- 分类是数据挖掘技术中运用最为广泛也是比较重要的分析手段，它是指运用训练数据集，通过分析数据的特征和运用一定的算法求得分类规则，该分类规则就是数据分类的模型，然后运用该模型对任何位置的数据对象进行分类。分类分为两个阶段：①构建分类模型，通过一定的算法对已知类标记的数据集建立分类模型；②用第一阶段构造的模型来预测给定的数据对象的类别。比较典型的分类方法有决策树分类方法、神经网络分类法、贝叶斯分类方法以及 K-近邻分类方法。

本章作业

1. 列举大数据处理过程包括哪几个步骤。
2. 简要说明数据采集的 4 个过程。
3. 简要介绍数据预处理的 3 种主要方法。
4. 简要介绍数据存储的 3 种典型存储方案。
5. 列举数据的 3 种主要来源。
6. 简要说明核心数据出现的最大问题。
7. 陈述外围数据的基本准则。
8. 陈述 HDFS 系统的特性。
9. 简要说明 HDFS 的结构及 3 种主要角色。
10. 列举说明数据挖掘中分类分析的主要方法。

第 3 章

大数据在商业银行中的应用

本章目标

- 了解大数据在商业银行客户关系管理中的具体应用
- 掌握客户生命周期管理的概念，了解大数据在商业银行精准营销中的具体应用
- 了解大数据在商业银行信贷管理中的具体应用
- 了解大数据风险控制与传统风险控制的区别，以及大数据在商业银行风险管理中的具体应用
- 了解大数据是如何帮助商业银行进行运营优化的

本章简介

与其他行业相比，商业银行在大数据技术的应用中具有独特的优势。这一优势主要来源于 3 个方面：首先，商业银行的业务系统信息化程度高，数据资源充足；其次，商业银行的数据规模庞大，数据种类较为齐全；再次，由于商业银行受到严格的监管，其数据的格式较为规范，数据的准确性也相对较高。因此，大数据在商业银行的客户关系管理、精准营销、信贷管理、风险管理、运营优化等方面中有着广泛的应用。

本章重点讲解大数据在商业银行客户关系管理、精准营销、信贷管理、风险管理和运营优化中的具体应用。





3.1 客户关系管理

客户关系管理这一概念起源于 1980 年年初在美国所出现的“接触管理”，之后由 Gartner Group 公司正式提出。Gartner Group 公司将客户关系管理定义为公司为了增加收入、增强盈利能力和提高客户满意度而提出的公司战略。具体来讲，客户关系管理包括两个层面的含义：一是公司要通过一定方式了解现有客户和潜在客户的需求；二是公司通过整合各方面的信息，从而实现对客户完整、一致的了解，且该过程贯穿于公司识别、筛选、获取、发展和保持客户的全过程。

当今商业银行都以“以客户为中心”的经营理念开展业务，因而客户关系管理在银行同业竞争中扮演着重要的角色。良好的客户资源、高质量的客户群体以及出色的客户满意度和忠诚度，可以帮助公司在市场中占据有利的竞争地位。因此，客户是商业银行生存和发展的重要资源。

商业银行通过进行客户管理，可以通过更高效、周到、便捷的客户服务提升其业务流程管理能力，从而降低银行的运营成本。此外，基于数据分析的客户管理可以使银行最大限度地满足客户个性化的需求，从而提高客户对利润的贡献度，实现客户价值的最大化。

客户关系管理是基于数据分析技术所进行的客户管理活动，能够通过数据的集成、挖掘和分析技术为企业的客户服务、销售决策提供自动化的解决方案。因此，客户管理活动需要公司不断提高其经营管理水平，进而促进其管理效率的不断提升。

在大数据的应用背景下，商业银行可以通过利用大数据分析技术所进行的客户管理提高其负债业务水平，规避贷款业务和中间业务中的风险。此外，虽然客户管理概念早已被我国商业银行所接受，但在实际的实施过程中仍存在着形式大于实际的问题。随着大数据技术在商业银行领域的应用，各个层次客户的金融需求、每个客户的个性化需求都将会得到极大的满足。

3.1.1 客户细分

1. 利用大数据进行客户细分的优势

客户细分又称为客户分类，是指将庞大的客户群体根据各种指标划分为众多细分的客户群。同一客户群中的客户具有相同或类似的特征，不同的客户群之间存在显著的差异和不同。

商业银行作为直接向社会公众提供各种金融服务的机构，其客户群体庞大且覆盖了所有层次的人群。因此在长期的金融服务中，商业银行积累了大量的信息数据，这些数据涵盖了客户的个人基本资料、收入情况、生活方式以及过往接受金融服务的历史记录等相关资料。商业银行通过利用先进的数据库系统和大数据挖掘及分析技术，对其所掌握的客户信息进行充分的利用，进而实现多个维度的客户细分。

1) 有效地维护和发展客户

在利用大数据技术进行客户细分的基础上，商业银行能够及时有效地获取不同层次现

有客户和潜在客户的需求、业务机会、相关成本及风险，并及时准确地制定相应的业务策略，从而向各个层次的客户提供个性化的服务和与其金融需求相匹配的业务推荐，使其各客户群都得到良好的维护和发展。

2) 运作效率的提升

在利用大数据技术进行客户细分的基础上，商业银行的运作效率也会得到提升。一方面，通过利用大数据技术进行的客户细分是更为有效的，商业银行可以在各个细分市场中发现新的业务机会，并及时采取行动把握发展时机，从而获得更多盈利。另一方面，在传统的商业银行客户关系管理中，存在着各信息系统相互独立的现象，即每个部门都有自己的客户关系管理系统，各部门间的数据无法实现共享，存在资源的浪费。而应用大数据技术，可以在同一系统中整合各部门的客户信息，并向各部门提供更加充分且多元化的信息资源。

3) 提高综合服务水平

利用大数据技术，可以对客户的相关资料和信息进行聚类分析，发现各个客户群的客户之间所存在的群体性行为，从而将这些具有同一共性特征但具有不同需求的客户组合成一个更大的新客户群。商业银行可以利用新客户群的共性特征，对他们在接受金融服务中的相似性进行把握，了解他们的投融资需求，进而提供有针对性的个性化服务，引导客户的投融资行为。在这一过程中，商业银行在降低服务成本的同时能够获取更高的收益，使其综合服务水平得以提高。

2. 客户细分的类型

1) 根据客户的风险和价值进行细分

这里所指客户的风险主要包括客户的信用风险和流失风险；而客户的价值即客户的利润贡献，可以通过利润率、营业收入等指标体现。根据客户的风险和价值进行细分，是通过对客户存款、贷款，以及其在理财产品、基金、保险等相关领域的金融活动进行辨识，分析客户为银行带来利润的主要业务以及相应的利润贡献水平。在此基础上，结合对客户潜在风险的分析和判断，将风险水平和贡献程度相当的客户划分为同一客户群。从中我们可以看出，商业银行根据客户的风险和价值进行客户细分，是基于其在客户关系管理中的投入和产出进行的，有助于提高商业银行的客户管理的有效性。

2) 根据客户交易行为特征进行细分

这里所指的客户交易行为，主要是指客户在进行金融活动时的交易金额、交易频率、交易对手等交易信息。在大数据技术的应用下，客户在进行金融活动时所产生的部分文字信息也可以作为客户的行为特征用以分析。例如，根据客户行为进行客户细分后，我们可以找到一类每月均匀发生多笔汇款业务且汇出大额款项的客户，对这一客户群有针对性地推出汇款费率优惠政策，以增强现有客户黏性，并吸引更多的同类型客户。

3) 根据客户的人口统计属性和行为偏好进行细分

客户的人口属性包括其年龄层次、收入水平等个人基本信息；而客户的行为偏好主要是指客户在其日常生活活动中所表现出的兴趣爱好以及生活方式，这一部分的信息主要是商业银行通过分析客户在使用其银行账户进行日常消费时的消费类目获取的。例如，根据

客户的年龄层次，可以判断出其在现阶段的金融需求：某一客户群体的年龄层次为 20～30 岁，且最近发生了一笔个人住房抵押贷款，商业银行可以及时地向该客户群体提供与其购买力相匹配的房屋装修信息，推荐住房装修分期业务，从而为客户提供恰当的个性化服务。图 3.1 列示了与客户相关的各类数据。

基本信息 · 姓名、年龄、证件号码、账户日均余额、卡内余额等	存贷款信息 · 存取款月均次数、存取款月均额度、贷款余额、贷款月还款额等	转账信息 · 转账月均额度、月均转账次数、转账备注信息、月均代缴费金额等
交易数据 · 月均交易金额、异地存取款次数及金额、业务类型、活动区域等	消费数据 · 月均消费金额、常用消费方式、购买产品范围、消费时间等	理财方式 · 是否有购买股票、基金、理财产品、保险、贵金属的行为及投资额度等

图 3.1　与客户相关的各类数据

3.1.2　预见客户流失

1. 捕捉流失客户的行为特征

随着金融市场竞争的日趋激烈，商业银行都在努力通过提供适时且多样化的金融服务吸引更多的新客户，并与原有客户建立良好的客户关系，以降低客户的流失率，从而使其获得长期利益。客户流失的发生具有明显的因果关系特征，而这些导致客户流失发生的原因通常可以在客户的账户状态、历史交易信息、服务反馈等相关数据资料中体现出来。

客户的账户状态、历史交易信息、服务反馈等数据信息通常是复杂且形式各异的，因此用传统的数据挖掘技术对这些信息进行分析是独立且效率低下的。而大数据技术在很大程度上弥补了传统数据挖掘技术的这一弊端，能够以高效的处理分析能力对上述数据信息进行处理，帮助分析人员得出及时有效的分析结果。因而在大数据技术的应用下，商业银行可以及时发现客户尚未被满足的需要和对现有服务的不满，及时采取恰当的行动解决客户的诉求，从而在客户结束其与银行的业务关系之前，及时对客户进行挽留，最大限度地减少客户的流失。

2. 对客户流失进行预测

商业银行应用大数据技术可以对客户流失进行预测。在客户关系维护中应用大数据技术，是从多角度对客户状态进行分析。因此，通过对客户流失的原因进行分析，构建出客户流失的预警模型，还能对潜在的客户的流失进行量化预测。在找出客户流失原因的基础上，根据客户相关信息与客户流失的内在联系可以构造出客户流失的关键性指标组合。从而使商业银行在日常运营中，通过利用实时监控所获得的相关指标数据预测客户的流失概率。通常采用决策树算法对流失客户的特征进行分析，从中获得流失客户和潜在流失客户的相关数据，并及时地将流失概率高的客户数据及时分配给客户服务部门。进而将预测结果与银行的促销手段相结合，实现客户忠诚度的有效提高。

3.1.3 高效渠道管理

1. 整合现有客户关系渠道

随着社会的飞速发展，商业银行中相互孤立的客户关系渠道已不再适应其在业务拓展、客户维护中的现实需要，因而商业银行需要对其各种客户关系渠道进行整合，建立起统一的客户关系管理体系。在原有的商业银行客户关系管理中，各渠道之间的信息存在严重的割裂，各渠道之间的数据缺乏有效的整合。例如，客户在柜台上办理了个人住房抵押贷款业务，而要想在微信银行上进行查询时，通常很难获得想知道的具体信息。

应用大数据技术可以通过对各渠道的客户信息进行采集，对客户的交易行为、消费偏好、潜在需求、忠诚度、社交关系等相关数据实现统一且有效的分析和整合，能够科学高效地对这些数据进行及时的处理，在客户从任一渠道接入银行系统时，该客户的所有相关信息都能及时反馈给相应的渠道。从中客户可以获得良好的服务体验。

2. 提高渠道管理的实时性

在当今的大数据时代，客户关系管理的渠道越来越强调及时性和敏感性。在当前客户关系渠道的多元化和社交媒介化的趋势下，商业银行通过利用大数据技术，能够实现客户信息的在线采集和交易行为数据的及时处理。在此基础上借助社交网络技术，结合客户的历史交易数据，通过对相关渠道数据的整合，商业银行可以实现对客户关系的实时精准的维护。例如，根据客户在电子商务平台中的消费记录判断出该客户的消费偏好，通过实时采集客户当前所在区域内的优惠信息，实时向客户提供与其需求相匹配的服务推荐。在这一过程中，客户关系管理的效率得到了提高。

3.1.4 推出增值服务，提升客户忠诚度

1. 发现客户尚未被满足的服务需求

在当前的大数据时代背景下，增强客户服务体验的方式也越来越多样化。通过应用大数据技术，能够有效地对客户的交易行为、交互行为及情绪状态进行整合并加以分析，从而帮助商业银行对客户的兴趣偏好和其对金融服务的使用轨迹进行准确的预测。商业银行根据预测对产品进行更加符合客户需求的体验化设计，对服务环境和服务流程进行优化，能够使其客户忠诚度得到有效的提高。

2. 提供恰当的增值服务

在商业银行向客户提供增值服务的过程中，通过整合各个客户关系渠道中的客户意见和看法，应用大数据技术能够帮助其从基于这些客户反馈的分析中，发现客户尚未被满足的服务需求，从而有意识地完善和提高客户在商业银行各渠道中的服务体验。结合当前 O2O 服务模式，商业银行需要通过产品创新和服务优化将线上服务和线下服务进行有机整合，结合客户和历史交易行为和当前密切关注的事件，提供个性化的增值服务。

例如，商业银行应用定位技术，根据客户所在位置和区域的不同及时向客户推送其附

近的各类商家给予本行客户的相关优惠信息，从而在客户心中树立高效、贴心的服务形象。有的商业银行还会基于各种客户关系渠道，向其客户定期开展理财知识讲解、新型诈骗的提示和防范、时尚潮流信息发布等各类服务信息。这些都是商业银行基于大数据技术向其客户所提供的增值服务。

3.1.5 案例——大数据帮助商业银行改善与客户的关系

1. 西太平洋银行集团

与客户共同发展成长是澳洲四大行之一——西太平洋银行集团(Westpac)一直遵循的价值观。随着数据源的增长和客户互动次数的增加，Westpac 开始了新的营销探索，他们将数据视为业务的血液。

在过去的两年多时间里，Westpac 借助 SAS 的分析工具打造了名为 KnowMe 的数据驱动营销平台，重塑与 1000 万客户的关系。2014 年，Westpac 每月会与客户进行 6000 万次来自网点、呼叫中心、ATM、移动端等渠道的互动。利用这些数据，Westpac 更加深入地理解客户需求，适时推荐客户正好需要的产品和服务。营销方式从“以产品为中心”向“以客户为中心”转变。这种转变也获得了市场和客户的认可，Westpac 的客户满意度高居澳大利亚银行业第一。

2. 法国兴业银行

法国兴业银行零售业务部门决策与研究经理 Joseph Emmanuel Trojman 确认说：“权力关系已经改变了，如今客户已经习惯于让银行之间相互竞争，而他们则等着更加个性化、更加及时的服务。”同时他还指出，最近一份 Cisco 集团的调研报告显示：近四成客户表示如果他们现在的银行不准备提供个性化建议，他们将选择更换银行。然而在法国，银行普及率已近 98%，对银行来说，找到新的客户是非常困难的。因此，法国银行界在意识到客户流失风险后，尤其是当这些客户的数据资产会给他们带来更丰厚的盈利时，对大数据的兴趣越来越浓厚。

在竞争异常激烈的个人储蓄市场，为了保持和增长市场份额，兴业银行最近分析了他们在法国 800 万个人客户的收入、储蓄等数据，并着重研究了公司分红、奖金和第 13 个月工资的发放日期。目的是为每一位客户确定推荐储蓄产品的最佳时机。Joseph Emmanuel Trojman 感叹说：“给 800 万客户做分析，工作量让人叹为观止。如果没有大数据的计算能力，电脑一定急得跳脚了。”其他的不说，只在处理时间上，大数据技术可以让之前需要三四天的数据分析缩短为三四分钟。

@ 3.2 精准营销

商业银行的精准营销是指在商业银行对其客户进行精准定位的基础上，结合不同客户的金融需求，依托信息技术手段，以向客户提供适宜的个性化服务推荐和产品营销。在精准营销的过程中，需要对客户的收入状况、业务类型、行为偏好、聚集区域、活动轨迹等

信息进行处理，从而使商业银行对其客户的认知程度、消费行为和金融需求形成科学的分析和预测(见图 3.2)。

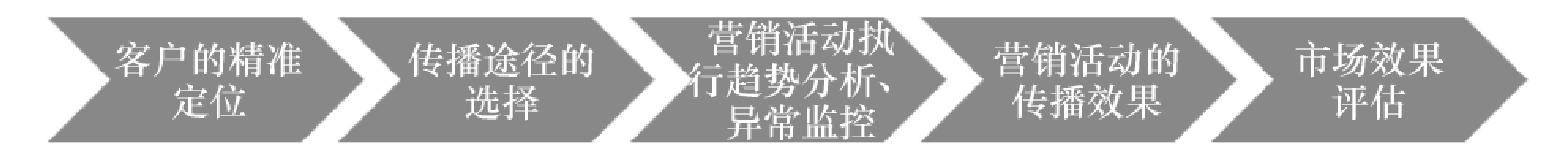


图 3.2　精准营销流程

上述通过客户行为分析并对客户需求进行预测以实现精准营销的过程离不开大数据技术的应用。例如，西班牙对外银行就推出了 ATM 机 ABIL，不仅在安全性和便利性上要优于传统的 ATM，而且还具备领先的记忆功能。客户在 ATM 上的历史取款金额和取款频率数据均被存储，在客户取款时 ABIL 会根据其存储的该客户账户情况给予客户取款建议。

传统的商业银行营销策略通常是基于对未来一段时间内的经济环境、监管政策、自身规模、客户资源、同业竞争等方面的考虑制定出来的，强调规模优势和实体经营优势而忽视客户的个性化需求。而在大数据时代，通过对多种渠道所获得的信息进行整合、分析和应用，进而满足客户的个性化需求才是商业银行成功营销的关键。

3.2.1　客户生命周期管理

1. 什么是客户生命周期管理

客户生命周期是指从企业尝试接触客户或客户开始了解企业时开始，到客户终止其所接受企业提供服务时结束的这一段时间。客户生命周期是产品生命周期的演变，对客户生命周期进行管理，有助于商业银行针对不同阶段客户群的需求特征，适时地采取有针对性的营销策略，从而实现精准营销以延长客户的生命周期。

客户生命周期可以分为客户获取、客户提升、客户成熟、客户衰退、客户流失 5 个阶段。在生命周期的不同阶段所要考虑的问题不同，应当采取的应对策略也不同，如图 3.3 所示。

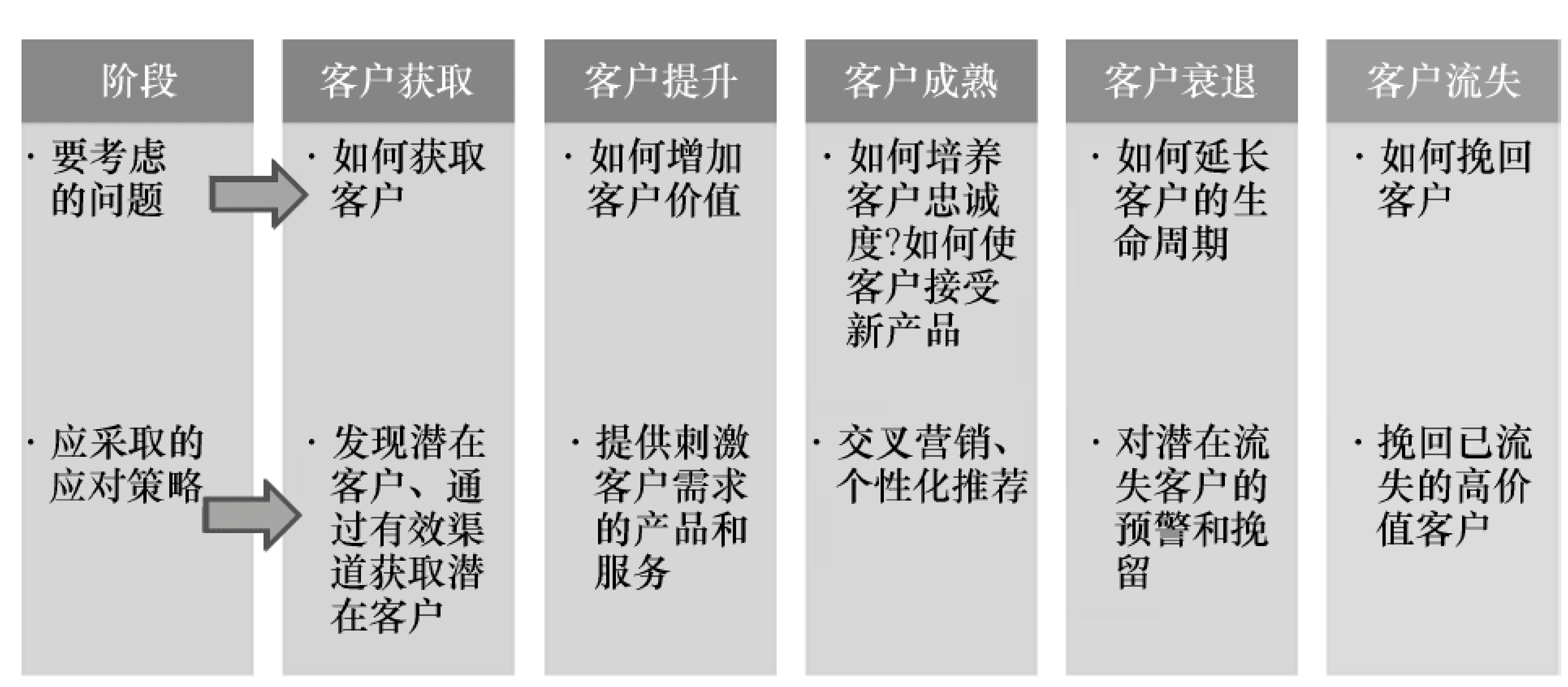


图 3.3　客户生命周期各阶段

2. 大数据技术在客户生命周期管理中的应用

商业银行应用客户生命周期管理的流程包括如下 4 个步骤。

(1) 从客户和市场的数据资料中挖掘客户尚未被满足的金融需求和市场发展趋势。

(2) 分析客户的交易行为、消费行为和兴趣偏好，以及客户为商业银行创造利润的价值驱动因素。

(3) 基于相关市场条件的假定，提出恰当的营销举措和政策，推出具有吸引力的产品和服务，以实现客户价值的最大化。

(4) 对新营销举措和政策、新产品和服务的市场反应进行追踪，及时改进相关营销策略和产品服务，在获得良性的市场反应后进一步推广。

在上述商业银行客户关系管理的流程中，大数据技术主要在步骤(1)和步骤(2)中发挥作用。①在客户获取阶段，通过利用大数据技术对新进客户的主要特征及关键购买因素进行分析，从而发现潜在客户群并选择有效的营销渠道来获取潜在客户。②在客户提升阶段，通过利用大数据技术分析现有客户的业务使用情况和主要行为特征，了解真实的客户反馈，进而发现基于客户需求的潜在市场空间以及客户价值提升的障碍，适时地推出满足客户潜在需求的新产品以及适应各类客户群的个性化服务。③在客户成熟阶段，通过利用大数据技术分析和跟踪成熟客户的深度需求和忠诚程度，进而对客户进行交叉营销和个性化推荐，以提高客户的黏性。④在客户衰退阶段，通过利用大数据技术分析和监控客户账户状态的变化，发现客户流失的主要驱动因素并对客户流失进行预测，在充分了解市场竞争态势的基础上，通过采取有针对性的营销举措最大限度地降低客户流失的可能性。⑤在客户流失阶段，通过利用大数据技术对流失客户的相关数据进行分析，找出客户流失的主要原因，进而采取有针对性的营销策略来挽回已流失客户。

3.2.2 实时营销

实时营销是指根据特定客户当前的个性化需要，向其提供商品或服务，并在客户使用该商品或服务时自动收集客户的使用信息，并对这些信息进行分析以了解客户的行为偏好和具体需求，进而自动对其产品或服务进行调整，实现对客户需求适应的实时性。实时营销是在传统营销的基础上发展而来的。虽然实时营销与传统营销都是以客户需求为出发点和主要着眼点，但传统营销所强调的仅仅是客户当前的需求，而实时营销与之相比更加强调客户的动态性需求，包括客户当前的需求和未来的需求。

1. 实时营销的特征

1) 满足客户当前的个性化需求

即在营销过程中，向客户提供的产品或服务要适应客户多种多样的个性化需求。为实现这一目标，商业银行在向客户进行营销前需要利用大数据技术采集和分析客户在使用产品或服务前所存在的现有需求，从而实现有效的营销。

2) 在动态过程中满足客户未来的需求

即在客户使用产品或服务的过程中，及时地获取客户每一时点的需求，从而通过完善其向客户提供的产品或服务，实现对客户动态需求的满足。商业银行通过利用大数据技

术，及时获取和分析客户在使用产品或服务中的需求的变化，从而及时地对其产品或服务的性能进行完善和修补，从而实现在动态过程中的有效营销。

3) “客户——产品”层的信息反馈模式

在传统营销中，客户信息的反馈处于“客户——公司”层，即客户需求的反馈信息是先传递给公司，公司基于反馈再对产品或服务进行改进和完善，从而再对客户的动态需求进行满足，公司是中心组织。而在动态营销中，客户的反馈信息处于“客户——产品”层，即产品与客户之间形成独立的关系系统，客户的动态需求能够及时地被产品接受并加以满足。商业银行的运营效率无疑会得到提升。

4) 适应的过程在无意识的状态下完成

在实时营销的过程中，客户是无意识地做出反馈的：客户无须专门提出意见、建议或完善资料信息；公司是无意识地做出产品或服务调整的：基于大数据技术的信息捕获和分析，产品或服务能够及时地对所发现的不足和空间进行调整和满足，无须公司层做出反应。

2. 大数据技术与实时营销过程

1) 感知阶段

在实时营销过程中，商业银行利用大数据技术，能够实时地对与客户产品使用和服务体验相关的电子记录进行获取和挖掘，从中获取客户体验的实时信息，从而有效地感知客户对其所提供的产品或服务的现实需求。

2) 分析阶段

基于感知阶段所获取的信息，大数据技术能够对这些信息进行自动的分析，从而帮助商业银行了解其所提供的产品或服务中所存在的主要不足以及发展空间。在运用大数据技术的基础上，商业银行对客户需求的获取在提高准确性的同时成本也会大幅降低，使商业银行对市场的把握能力得以增强。

3) 适应阶段

基于在前述两个阶段运用大数据技术，商业银行对市场的感知能力和分析能力已大幅提高，此时产品或服务的适应能力就成为实时营销的关键。因此，商业银行在对其产品或服务进行设计时，应当在其中使用一定成熟的智能技术，从而帮助其产品或服务对客户需求的动态变化做出及时有效的调整。

3.2.3 交叉营销

交叉营销就是基于所发现的客户的多种需求，通过销售多种相关的产品或服务来满足客户需求的营销方式。换言之，交叉营销是一种从横向角度开发市场的营销方式，在这一过程中客户的多种需求能够同时被发现和满足。

1. 银行业中的交叉营销

交叉营销在银行业的作用尤为明显。因为客户在购买银行所提供的金融产品和服务时需要提交一定的个人资料，其购买行为也会被记录下来形成电子资料，这些数据资料可以

帮助商业银行分析和了解客户需求，从而为其客户提供更多更优质的金融产品和服务。此外，这些数据还可以在保护客户隐私的基础上，与商业银行的互补金融企业之间共享，进而实现互助营销。

商业银行在当下所面临的主要挑战不再是市场份额的竞争，而是利润份额的竞争，因而商业银行的着眼点不再是一味地扩大规模，而是努力提高每个客户的贡献程度。采用交叉营销的策略，能够帮助商业银行以最低的成本使客户尽可能同时拥有多种银行所提供的金融产品或服务，进而使银行的利润得以增加，使其客户忠诚度也得以提高。

2. 大数据技术在交叉营销中的作用

交叉营销成功的关键在于：找对人、说对话和做对事。基于现有的客户数据资料，商业银行借助大数据技术可以对其所掌握的客户资料进行整合和关联性分析，进而高效地发掘出客户潜在的多样且相互关联的需求，进而有针对性地进行交叉销售。

1) 找对人

找对人是指要找准具体的客户群体。商业银行利用大数据技术可以对客户使用金融产品和服务时的行为特征进行分析，并根据这些特征将客户分成组内特征相似、组间特征不同的群组，进而发现针对不同的客户群的市场机会。

2) 说对话

说对话是指通过对客户数据进行分析，选择有效的促销渠道。商业银行通过大数据技术还可以了解到不同客户群的心理特征和行为偏好，进而使商业银行在找对人的基础上，能够根据不同客户群的偏好对其不同的目标客户进行有针对性的宣传和营销活动。

3) 做对事

做对事是指向目标客户推荐与其需求相符的产品或服务。商业银行利用大数据技术能够发现产品与产品之间、服务与服务之间、产品与服务之间的关联规则，找出最优的产品或服务组合，进而提高商业银行在找寻组合销售机会时的准确性。

3.2.4 社交化营销

社交是指人们之间传递信息、交流思想的交际往来活动。社交化营销即企业有意识地利用社交活动进行营销。随着移动互联网的不断发展，越来越多的商业银行开始重视运用网络手段创造价值、提高品牌影响力，以期在当前激烈的市场竞争中出奇制胜。

1. 商业银行进行社交化营销的动因

1) 客户消费行为的演变

移动互联网的出现和社交媒体的普及代表着人们在新时代行为方式的转化。过去人们使用电脑浏览网页、在线支付、汇款转账，但现在越来越多的资金划转和收付是在移动终端设备上完成的。因此，为了避免出现商业银行与客户之间的隔绝，商业银行需要顺应时代潮流，采取新的营销方式接近客户。

2) 增进与客户之间的联系与互动

现在人与人之间的联系和沟通越来越多地依赖社交媒体。商业银行主动与社交媒体相

结合，无疑可以增进其与客户之间的联系，增强客户的服务体验。商业银行可以利用社交媒体与客户进行互动，收集客户对于其产品的问题、评价、反馈和建议，进而拉近与客户之间的关系、对其所提供的产品和服务进行完善和改进。

3) 信息传递速度快、范围广、针对性强

社交媒体作为人与人之间沟通的媒介，具有直接、快速和便捷的特点。因此，商业银行借助社交媒体进行社交化营销，可以将产品和服务信息直接传递到客户手中。而且由于社交媒体的用户数量庞大，通过社交媒体进行营销有着广大的受众范围。此外，借助大数据技术和定位技术，商业银行可以向客户提供与其需求特征相符的针对性营销。

4) 获取客户信息的能力增强

商业银行借助社交媒体可以与客户进行直接的沟通，也可以通过收集客户基于社交媒体与银行所发生互动的相关数据并加以整合和分析，对客户需求、现有产品和服务的市场反应和不足形成清晰的了解，从而使商业银行的营销更加高效。

2. 大数据技术在社交化营销中的运用

1) 获取信息

在社交化营销中，商业银行利用大数据技术可以基于其在社交媒体的后台直接获取多种多样的数据信息，这些数据信息是后续客户需求分析、产品与市场间关系把握和具体营销策略制定的基础。

2) 分析需求

在信息获取的基础上，商业银行利用大数据技术对其从社交媒体平台上所获取的信息进行分析和挖掘，进而使其对客户的多样化需求、现有产品和服务的不足以及市场空间形成清晰的认知。

3) 高效营销

在前述工作的基础上，商业银行将分析结果与市场趋势相结合，能够制定出与不同客户需求相适应的针对性营销策略，从而使其营销效率得到大幅提高。

3.2.5 个性化推荐

个性化推荐是指根据客户的交易特征和行为偏好，向客户推荐其可能感兴趣的产品、服务和信息，从而实现交叉销售的营销行为。个性化推荐的实现离不开大数据技术的运用。

随着社会的不断进步和发展，商业银行所提供的金融产品和服务越来越多，客户在基于自身的金融需求对其进行筛选时，难免要花费大量的时间和精力。而在这存在信息过载问题的过程中，很可能会导致客户的流失。为解决这一问题，商业银行可以运用大数据技术建立面向其客户的个性化推荐系统。

在个性化推荐系统中，大数据技术被用于对商业银行从各个渠道(例如，跟踪客户的浏览购买信息)所获取的海量客户数据进行充分的整合挖掘，从而帮助商业银行为其客户提供个性化的金融决策支持和信息服务。在这一过程中，商业银行不再是根据客户的关注和浏览数据就进行产品或服务的推荐和营销，而是在对客户数据进行纵向分析的基础上，对客

户的行为偏好进行充分的分析和挖掘，从而找出客户的共性行为并加以推荐。商业银行通过利用大数据技术分析一段时间内客户行为与产品和服务的关联性、金融产品和服务的购买频率和偏好，能够根据模型的自动学习演变功能预测出客户未来的购买需求和购买时间，从而及时地向客户做出适当的推荐。

3.3 信贷管理

信贷管理是指商业银行在国家现行的法律规定和相关政策的约束和规范下，根据安全性、流动性和收益性的原则对其所发放的贷款进行贷前调查、贷时审查和贷后管理的过程。商业银行信贷管理的目标是降低其信贷业务的风险，以实现信贷业务的效益最大化。

3.3.1 贷款风险评估

贷款是商业银行最为主要的资产，也是影响其经营能力的关键因素。因此，商业银行要最大限度地降低和控制贷款风险，对贷款风险进行评估是必要的。

1. 传统贷款风险评估所面临的挑战

具有快捷、简化和纯信用特点的互联网金融正在快速发展，给传统的信用风险评估带来了不小的挑战。其中，快捷性主要体现在贷款申请和审批的快速和便捷方面，要求贷款审批过程要实现自动化和系统化、减少人工审批所占比重。简化性和纯信用性则主要体现在客户申请贷款时所需提交材料数量的减少和申请过程的便利方面，但简化性可能会导致更加严重的信息不对称，纯信用则对商业银行客户风险评估的准确性提出了更高的要求。

从中我们可以看到，商业银行在创新模式下进行贷款风险评估需要从多个维度获取客户信息，并利用有效的风险计量技术对其所面临的贷款风险进行合理评估。

2. 大数据应用下的贷款风险评估

1) 信息输入的多样性

传统的贷款风险计量主要是利用贷款申请人的申请信息、中国人民银行的征信信息建立信用评分模型和风险规则的。利用大数据技术对风险进行计量突破了传统计量方法的限制，不再仅依靠传统数据对风险进行计量，而是将更多的非传统数据纳入风险评估系统，从而可以更全面地对贷款人的信用状况和风险程度进行评估。从中可以看出，商业银行运用大数据技术对贷款风险进行分析评估，能够帮助其在保证评估结果准确性的基础上，优化贷款审批流程，提高贷款申请和审批的速度和便利程度，进而使商业银行的经营效率和同业竞争力得以提高。

2) 评估过程的自动化

在大数据背景下的贷款风险评估过程中，实现了授信审批的流水线作业，并呈现出自动化的特点。如图 3.4 所示，贷款风险评估通常包括获取申请信息、外部信息采集、决策规则校验、电话核实、评分卡运行、得出审批结果这 6 个环节。这 6 个环节的先后顺序是可以根据实际情况进行有机调节的，也可以根据实际情况的需要改变其串并联方式。

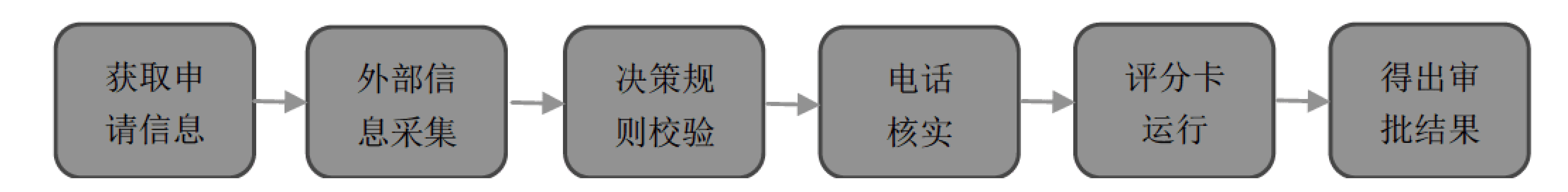


图 3.4　贷款风险评估流程

例如，对于资信情况较差的客户群，由于进行外部信息查询时商业银行需要支付相应的查询成本，而该类客户的审批通过率较低，对其进行外部信息查询无疑会增加运营成本、降低资信评估系统的运行效率。因此，可以针对该类客户调整贷款风险评估过程，如图 3.5 所示。

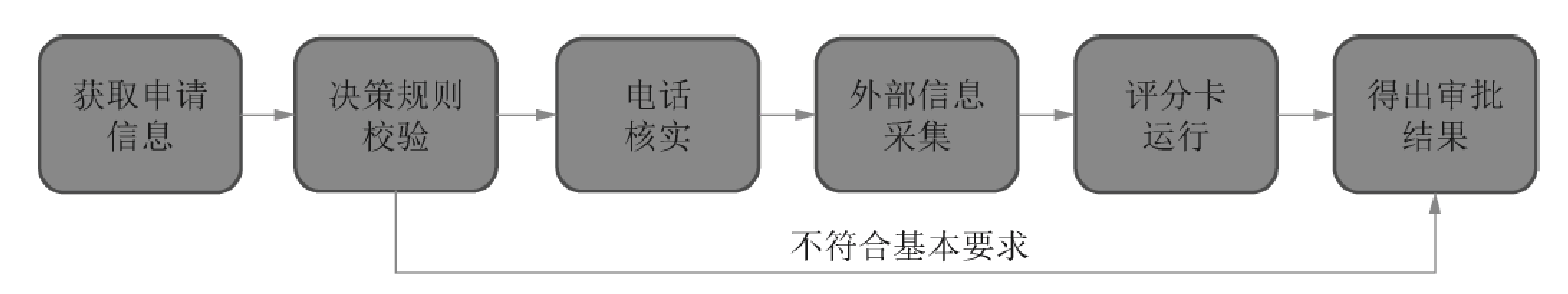


图 3.5　贷款风险评估流程——变形一

此外，为提升贷款风险评估过程的处理效率，可以将外部信息采集、决策规则校验和电话核实 3 个环节同时进行，在得到 3 个环节的结果后直接进入评分卡运行环节，如图 3.6 所示。

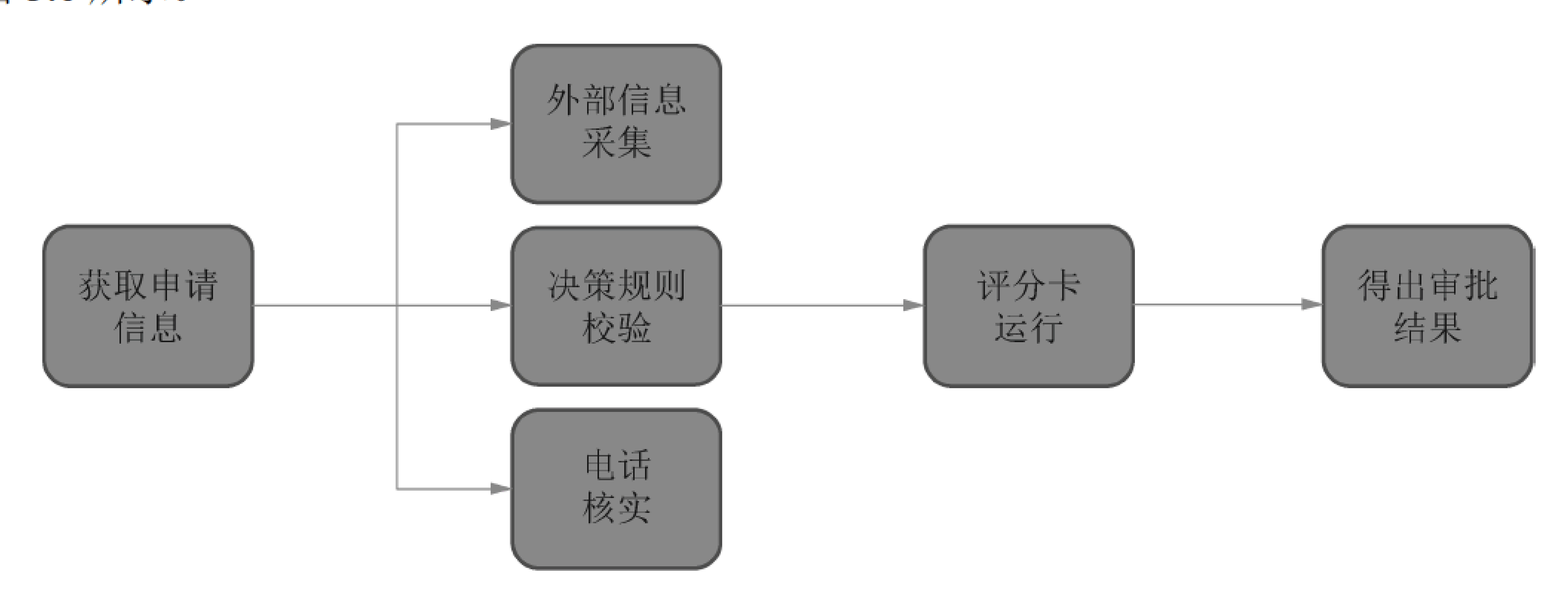


图 3.6　贷款风险评估流程——变形二

3. 客户风险评估模型

贷款申请风险模型是商业银行在信贷管理中最为常用也是最为重要的模型。该模型是在对客户多方面的信息数据进行分析挖掘的基础上，形成对客户资信状况的综合评价，从而判断向该客户提供信贷服务的风险。该模型除了可以帮助商业银行准确地识别和引入优质客户外，还能帮助商业银行制定差异化的客户管理策略。

在该模型中，客户风险是被衡量对象；目标变量由客户的逾期情况决定，在考核期内逾期天数超过给定阈值即为劣质客户，未发生逾期或逾期天数未超过规定天数的为优质客户。预测变量则根据客户所提交的申请信息、客户过去的历史交易信息和客户在第三方的行为信息确定，包括但不限于以下几个方面：工作单位、家庭状况、资产负债情况、学历

层次、历史信贷行为、基本账户流水、公积金状况、社保状况、社交状况、网络交易行为、浏览行为等。

贷款风险评估模型主要应用于贷款申请阶段，服务于客户准入。对于经模型判定得分较低的客户，商业银行通常直接拒绝其贷款申请；对于那些评分在事先所设定好的准入评分阈值附近的客户，商业银行可以通过进行二次评估来决定是否向该客户授信；而对于评分较高的客户，商业银行可以直接做出授信决策。

在该模型中，传统的信息数据多为静态信息，对客户风险评估所起到的作用十分有限。而基于大数据技术所引入的多渠道行为数据多为动态信息，弥补了静态数据的不足。因而在贷款风险评估中将大数据与传统数据相结合，能够有效地提高模型的评估能力。

3.3.2 信用卡自动授信

商业银行传统信用卡授信方式是人工审核申请资料，然后根据客户大致的风险等级发放相应的授信额度或拒绝申请。在信用卡用户使用信用卡的过程中，商业银行积累了大量的信用卡客户数据，可以把是否违约、违约概率、有效使用额度等指标作为评价对象，然后调用与此相关的各种客户信息建立评估模型，自动计算授信结果。

信用卡产品是循环授信产品。在大数据技术的应用下，商业银行可以对信用卡客户的信贷风险进行实时监控，并根据监控结果及时对客户的授信额度做出调整。具体来讲，大数据技术可以主要应用于对信用卡客户的初始额度确定、业务风险评估、业务收益评估之中。

1. 初始额度模型

初始额度模型衡量的主要是商业银行基于其信用卡客户的收益情况。由于信用卡产品是需要循环授信的，因此商业银行在进行初始额度授信时，除了要考虑客户需求和还款能力之外，还会考虑客户的收益情况，进而其信贷资源会向高收益客户倾斜。

商业银行在信用卡业务中的主要收益来源于其向客户收取的循环利息、逾期利息、手续费等利息费用。因此，在初始额度模型中所考虑的变量主要包括：客户属性(性别、年龄、学历等)、还款行为、逾期行为、额度占用情况、透支情况以及客户在电商平台上的消费行为、分期行为、浏览行为、点击行为等与客户相关的数据信息。通常情况下，习惯分期的客户收益率较高。

大数据技术在初始额度模型中能够发挥重大作用。在客户的初始申请授信阶段，商业银行尚未与客户建立直接的业务关系，基于收入、负债等基本信息难以对客户的授信需求做出合理的评估。商业银行应用大数据技术可以通过外部客户的相关交易数据对客户的消费支出情况进行分析，从而能够较为准确地评估出客户的授信额度需求。

2. 行为风险模型

行为风险模型是根据客户历史行为预测其未来出现坏账的可能性，进而对客户风险做出全面准确评价的模型。该模型在商业银行的信用卡额度管理中起到了重要的作用。

在信用卡客户使用信用卡的过程中，商业银行能够观测到客户更多的行为。对这些行

为数据进行分析和挖掘能够为商业银行提供有效依据，以实现对客户所提供的循环授信更为准确有效的动态调整。

目标变量由客户的逾期情况决定，在考核期内逾期天数超过给定阈值即为劣质客户，未发生逾期或逾期天数未超过规定天数的为优质客户。这里阈值的确定是由多方面因素所决定的。例如，可以根据逾期天数和客户未来造成损失间的相关性，对劣质客户和优质客户进行区分。或根据客户数据的累计周期的长短，分别设定阈值标准：累计周期短，逾期天数标准就越低；累计周期越长，对劣质客户的判别就更依赖客户的行为数据分析而不再局限于逾期天数的情况。

在行为风险模型中，客户行为的预测变量通常包括但不限于以下方面：客户的还款行为、消费行为、资金使用情况、欠款情况、取现行为、银联流水数据、央行征信数据以及客户在电商平台上的消费行为、浏览行为等。

3. 业务收益模型

业务收益模型与初始额度模型类似，所衡量的都是客户能够为商业银行所带来的收益情况。但业务收益模型的衡量对象是商业银行的存量客户，主要用于对其存量客户的收益情况进行动态评估。商业银行的信贷资源通常会向低风险、高收益的客户倾斜，并缩减分配给高风险、低收益客户的信贷资源。从中我们可以看出，在商业银行的授信额度调整策略中，不仅会考虑客户的行为风险，还会考虑客户所能带来的收益。因此，将业务收益模型与行为风险模型相结合，有助于商业银行保持合理的资产结构。

综上所述，大数据对计量模型的影响主要表现为以下 3 个方面。

(1) 大数据提高了计量模型的信息完备性。

(2) 大数据提高了计量模型的精益化程度。

(3) 大数据使计量模型和业务决策的过程更加及时、结果更新更为频繁。

3.3.3 案例——大数据为商业银行信贷管理提供更多可能

1. 商业银行开始意识到大数据的重要作用

2012 年中国建设银行“善融商务”的率先上线，为大型商业银行涉足电商领域拉开了帷幕。随后，中国交通银行的“交博汇”、中国银行的“云购物”和中国工商银行的“融 e 购”纷至沓来。目前，几乎所有的银行系电商都声称，免平台费、免技术维护费、免交易佣金费。而这些“免费”的背后，是银行朝思暮想的数据信息。

为获得客户的真实数据，银行往往要进行大量线下调查工作，成本巨大。与电商合作只能得到信息的分析结果，但银行更希望自己做电商获得一手数据。电子商务平台上积累的大量数据，比如消费者的搜索、比价、商户流水等，能够转化为银行评级、授信的数据，会对银行发展潜在客户、规避信贷风险起到重要作用。

2. 商业银行纷纷推出小微企业大数据产品系列贷款

中国建设银行为资金周转困难、受限于贷款烦琐手续的小微企业主推出了小微企业大数据产品系列贷款。九大系列产品包括小微快贷、税易贷、善融贷、结算透、信用贷、创

业贷、POS 贷、薪金贷、善融 e 贷。大数据信贷产品是建设银行运用大数据技术，对小微企业客户的结算、交易、存款、资产数据、信用记录等信息进行分析判断、主动挖掘和营销发放的信用贷款，具有小额化、标准化、综合化、集约化、智能化的特点。

3. 大数据时代的机遇和挑战

伴随科技的发展，移动互联的浪潮汹涌而来，大数据、云计算等科技因素不仅可以支持银行提高效率、降低成本，持续激增的数据还能迫使银行寻求新的方法来采集、整理数据，助推金融创新。各大银行对大数据的争夺风生水起，然而过程中可能产生的数据造假问题也不容小觑。大数据固然能降低商业银行的交易成本，但也能降低客户的造假成本。例如，POS 贷本应基于“真实贸易交易”用户的信息，但在实际操作过程中，也可能存在个别商户为了提高贷款额度，进行“流水造假”的现象。

3.4 风险管理

风险管理是指企业在其日常经营活动中努力将风险降到最低的管理过程。在这一管理过程中，企业需要对其所面临的风险进行认识、度量和分析，通过科学决策选择出最为有效的风险管理途径和方法，力图通过具有主动性、目的性和计划性的风险防控行为，以最小的成本获取最大的安全保障。有效的风险管理活动能够帮助商业银行降低损失出现的概率、缩小损失的影响范围，进而提高其经营能力和市场价值。

近年来，互联网金融的迅猛发展给传统金融机构带了极大的挑战。商业银行在过去主要以中国人民银行所提供的征信信息和客户所提供的基础信息为主要的风控信息来源，以专家经验为风险管理决策的评判方法，过于倚重定性分析可能会错失部分有效客户，不利于其业务的开展。随着移动互联网的普及，人们的日常活动越来越多地在网络上留下痕迹，这些痕迹可以以电子数据的形式存储下来。有越来越多的商业银行开始运用大数据技术对客户行为进行获取和分析，以对其风险控制活动进行有效的补充。

商业银行所面临的风险包括：信用风险、操作风险、市场风险、流动性风险、利率风险、法律风险等。其中信用风险是商业银行所面临的最主要的风险。因此，下面主要针对信用风险管理进行阐述。

3.4.1 大数据风险控制与传统风险控制的区别

随着移动互联网时代的来临，人们在网络上所留下的行为印记越来越多，这些类型多种多样的印记作为数据被存储下来，已经成为金融机构金融风险控制的重要补充手段。运用大数据进行风险控制能够很好地弥补传统风险控制所存在的信息不对称、数据获取维度窄、人工采集成本高、效率低等缺点。

1. 大数据风险控制与传统风险控制间的差异

1) 传统风险控制

传统风险控制流程如图 3.7 所示。在用户提交申请表后，商业银行首先要查询客户的

征信情况；由录单员负责，将申请表中的客户信息录入系统，并另行登记审批进度表；之后将客户申请资料随征信资料派给审核员；审核员通过阅读征信资料、查询信用网、工商信息、与第三方核实申请资料和确认申请人真实性等审核步骤后，记录存在的疑点；电话联系客户，对审核中发现的疑点进行核实；之后对申请人进行实地考察，咨询其经营模式、营业收入等问题，对其经营场所、经营状况等信息进行核实；在贷款分析环节，结合之前进行的调查情况撰写调查报告，给出审批意见；进而结合审批意见，做出信贷决策；通知审核通过的客户来行进行签约，在签约的过程中要进行复核相关资料的原件、核实客户流水情况等流程；在放款给客户后，对相关文件进行归档；在客户借款期间，要做好贷后管理，包括电话回访、通知还款、催收、续贷等业务活动。从中可以看到，传统的风险控制流程十分烦琐，复杂的流程无疑会导致业务办理的低效率。

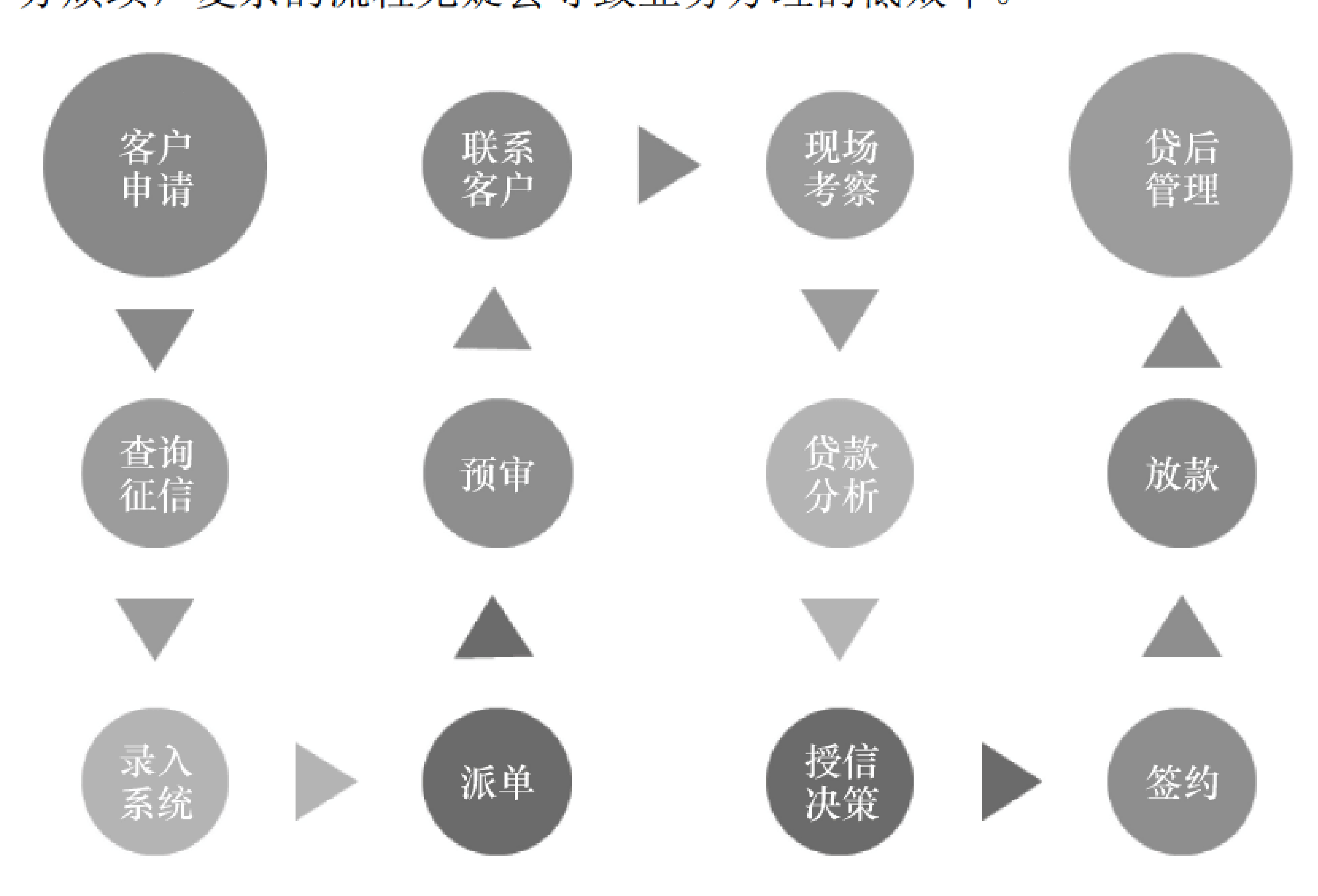


图 3.7 传统风险控制流程

2) 大数据风险控制

大数据风险控制流程如图 3.8 所示。具体来说，在大数据风险控制中，客户通常从网页端口或手机客户端口(这些端口也就是数据采集的入口)进入贷款申请系统；商业银行在获得客户授权指令后，利用其系统内和第三方的相关客户信息数据对客户进行征信：首先是对客户身份进行验证，并对其进行黑名单检查，之后利用客户的交易行为数据、社交数据、教育数据、运营商数据、电商数据、公积金数据、社保数据等相关数据对客户的信用风险进行分析和评估；在评估结果的基础之上，生成该客户的资信报告；基于资信报告做出授信决策，并向客户发放贷款；在客户借款期间，在与客户保持联系的基础上，依据事先设定好的催收模型和催收策略对客户的信用风险进行实时监控。从中可以看到，大数据风险控制的基本流程与传统风险控制大致相同，但在接受客户申请、对客户进行资信评估、做出授信决策、进行贷后管理环节比传统风险控制更加快捷高效。

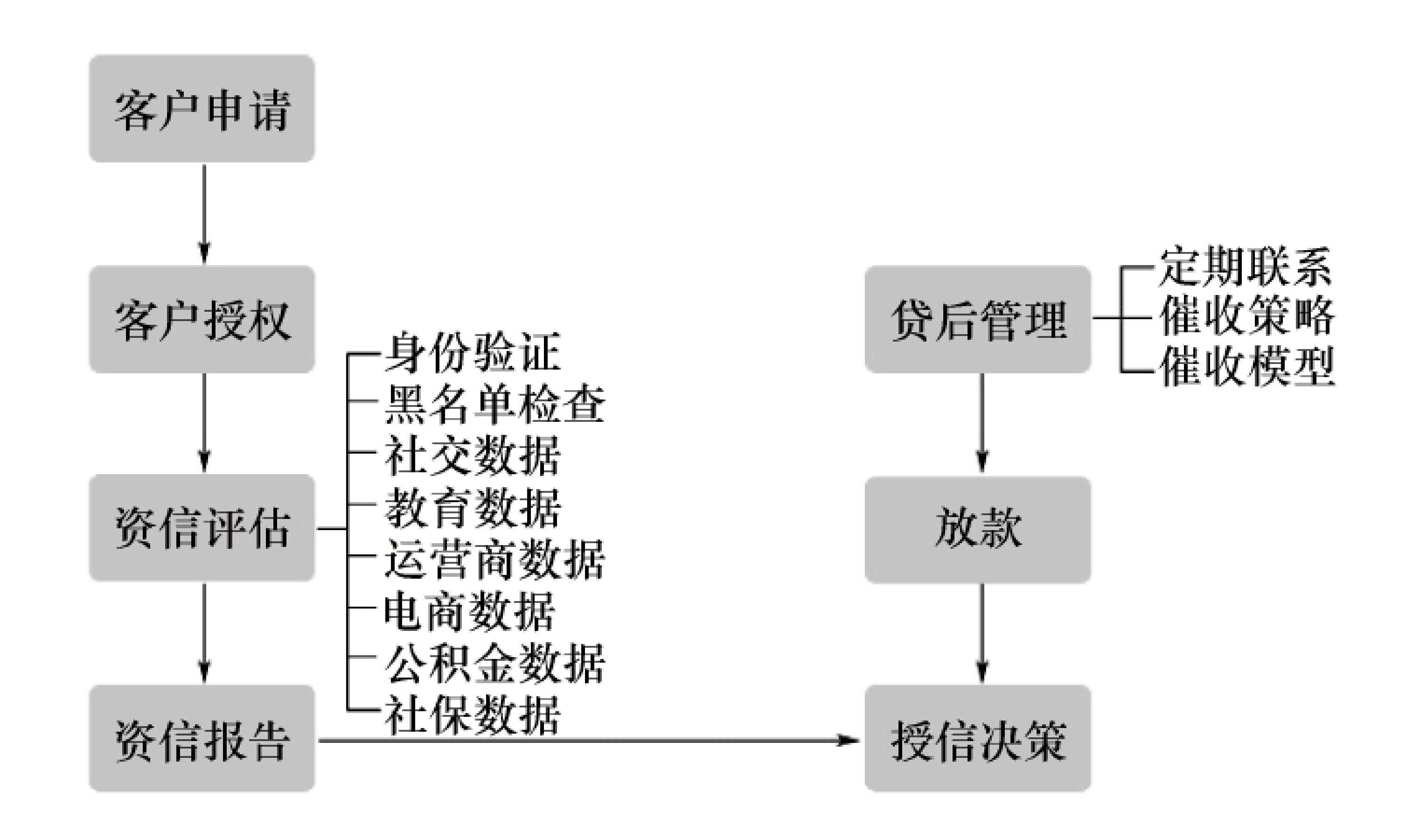


图 3.8　大数据风险控制流程

3)　二者之间的差异

大数据风险控制与传统风险控制最主要的差异体现在大数据技术在客户征信环节的运用。大数据征信与传统征信的不同主要体现在以下 6 个方面。

(1)　数据来源不同。传统征信的数据以银行信用数据为主，来源单一，采集的频率相对较低。而大数据征信的数据来源广泛，包括：用户提交的数据，如其职业背景、受教育程度等；第三方数据，如理财数据、电商平台数据、社交平台数据、社保数据、公积金数据等其他相关数据；此外，大数据征信的信息采集频率高，能够实现对数据的实时采集。

(2)　数据格式不同。传统征信所采用的数据主要是格式化数据；而大数据征信所采用的数据既包括格式化数据，也包括大量的非格式化数据。

(3)　评价思路不同。传统征信是通过客户历史信用记录来评价客户信用水平的；而大数据征信则不仅对客户的历史信用数据进行考量，还会从海量数据中推断客户的身份特质、性格偏好、经济能力等相对稳定的指标，从而对客户的信用水平做出判断。

(4)　分析方法不同。传统征信所采用的分析方法主要是线性回归、聚类分析和分类树等方法；而大数据征信所采用的是机器学习、神经网络、Page Rank 算法、RF 等大数据处理方法。

(5)　服务人群不同。传统征信的服务范围仅限于有信贷记录的客户，服务范围小；而大数据征信的服务范围不仅包括有信贷记录的人群，还包括那些没有信贷记录但在生活中留下足够多痕迹的客户，服务范围大幅拓展。

(6)　应用场景不同。传统征信通常只能应用于金融领域；而大数据征信不仅能应用于金融领域，还能在多种生活领域发挥其使用价值。

2. 大数据风险控制的优势

大数据风险控制的优势主要体现在大数据征信的利用价值上。

(1)　使商业银行的客户信用风险评估纳入了多样化的行为数据，这些数据覆盖范围广泛且具有实时性。依托于大数据和云计算技术的优势，可以对所收集到的海量数据进行充分挖掘，从而使商业银行的客户行为风险模型不断迭代优化。

(2) 在大数据风险控制中，信用评价更加精准。由于大数据征信模型中客户数据的范围越来越大、数据维度越来越广，客户信用评估模型越来越多，因而依据大数据征信模型所做出的信用评价更加精准和高效。

(3) 大数据风险控制中对客户信用的评判更具时效性。大数据所具备的数据采集和计算能力可以帮助商业银行基于多维度、全方面的客户数据以及具备自我学习能力的风险控制模型，获取实时计算出的评估结果，进而使其风险量化能力得以大幅提高。

3.4.2 基于大数据的银行风险管理模式

1. 基于大数据的银行风险管理模式所具有的特点

1) 集约化管理

在大数据技术的应用下，商业银行触及客户的方式发生了极大的变化，其在对客户信用风险进行管理时无须以现场直接接触的方式接触、服务和管理客户，而是以电话联系、网络在线沟通、移动智能设备客户端等方式与客户进行互动，进而有效地降低了运营成本。此外，由于业务流程更加标准化，在保证提高业务质量的同时，商业银行的服务效率也得到了提升，从而能够更好地在控制风险的基础上向不同的客户群提供其所需的金融服务。

2) 全过程风控

商业银行基于对大数据技术的应用，能够在其风险管理系统中接入海量集中式数据，这些多维度数据的交叉验证，能够解决商业银行在客户信用风险评估中客户信息难以收集的问题，从而有效地缓解了商业银行在信贷业务中所面临的信息不对称的问题，提高了商业银行对客户信用风险的识别和预防能力。

此外，基于对大数据技术的利用，商业银行的贷后管理能力也得到了提升，尤其是非现场的贷后管理能力得到了大幅提高。在大数据技术的应用下，商业银行的风险控制以非现场的预警监测为依托，对不同客户群的风险特征和行为模式进行识别，强调对授信客户进行持续跟踪、动态监测和实时预警。

3) 标准化与差异化相结合

虽然商业银行所提供的信贷产品具有一定的标准化特征，但在其风险管理过程中也同样会考虑如何对差异进行处理。根据数据分析和市场调研的结果，商业银行可以针对不同行业、不同地区、不同特征的客户群制定不同的标准化产品，并分别采用不同的运作流程、审核标准、评分卡和授信策略。在集约化的风险管理下，商业银行可以在不断的学习和测试过程中，对其经营策略进行细分和调整。

4) 输入信息多样化

在大数据技术的应用下，越来越多的外部信息也被纳入商业银行的风险评估系统。在对外部信息进行标准化处理后，信息数据之间所进行的交叉验证能够在结合各个客户群特征的基础上进行优化。随着外部输入信息的范围越来越广、数据量越来越大、数据变化频率越来越快以及数据类型愈加多样化，商业银行的风险管理系统在数据处理、数据分析、模型建立、策略应用等方面的能力也在不断增强。

2. 信贷审批

信贷审批是商业银行进行风险管理的重要环节。随着社会的不断发展和商业银行同业间的竞争加剧，商业银行在进行信贷审批时越来越注重客户的体验。例如，提供更加简便的贷款申请流程、更快速的审批结果反馈、更公开透明的贷款受理过程等都是提升客户审批体验的主要表现。在保证风险控制水平和能力的基础上，提升客户的审批体验离不开大数据技术的应用。

1) 实时审批

实时审批是自动化审批的一种类型，是指从获取申请信息开始，通过接入外部数据并进行比对、规则判断、信用调查和模型评估，到最终给出授信决策，在保证决策质量的前提下整个过程是在极短的时间内完成的。

为了实现实时审批，商业银行需要对其审批流程进行优化，减少人工干预的必要性，还需要对其非人工环节的运行效率进行提高。具体来讲，就是要让数据、模型和策略更多地代替人工做出判断，并对信息技术进行革新，以智能决策模型和策略进行操作。例如，在有效信息足够完备的情况下，利用第三方的数据信息就可以对客户的申请信息进行校验和补充，无须工作人员再电话联系客户核实信息的真实性和完整性。

大数据是实时审批的根本。在大数据技术的作用下，客户所提交的申请资料得以简化，使客户的审批体验得到了有效的提升。此外，商业银行基于大数据技术也不再单纯依靠客户所提交的信息对客户的信用风险进行评估，而通过分析其他渠道获取的真实数据所得出的评估结果无疑更为有效。

2) 前置审批

利用大数据技术，商业银行可以结合多个渠道的客户数据，在客户提交信贷申请前就对客户的风险水平做出评估，预先做出授信的决策，即将审批过程前置。如此一来，商业银行的工作人员根据审批合格的客户名单有针对性地接触这些优质客户，只要该客户提出授信申请便能直接与商业银行建立起信贷业务关系。从中可以看出，前置审批既是风险控制过程的一部分，也是营销环节的一部分。

大数据技术在前置审批过程中的作用表现为两个方面：一是能够使商业银行在对客户风险进行评估时使用到更加全面的数据，从而做出合理的授信决策；二是能够使商业银行对客户的信贷需求做出准确的预测，从而在恰当的时机为客户提供信贷服务。

3) 隐性审批

隐性审批主要存在于消费金融领域，即在客户进行消费付款时，及时为该客户提供消费贷款，无须客户专门提交授信申请。隐性审批过程有以下 3 个突出特点。

(1) 隐性审批有很强的应用场景。隐性审批通常与存在客户借款需求的应用需求相联系，发生于该客户在该场景中的付款过程之中。基于该应用场景，商业银行能够获取借款客户的资金用途信息，从而保证了信贷资金使用的真实性，是对客户资信状况的有效补充。

(2) 在这一过程中，授信申请、授信审批、放款和交易紧密地衔接在一起。即客户在发生交易行为时并未感受到其授信申请行为，授信审批和款项的拨付都集成在客户的支付行为当中。

(3) 维护商圈的过程就是寻找客户的过程。在隐性审批的过程中，商业银行只需要找到客户集中的商圈便可以轻松引入优质的借款客户。

大数据技术的优势作用主要体现在隐性审批时，商业银行对其借款客户的风险和收益水平的实时评估之中。利用更能反映客户消费能力和经营状况的第三方数据对客户进行评估，所得出的评估结果更加贴合客户的真实情况。依托于大数据的手机和存储，营销和审批环节更为紧密地结合在一起，使商业银行在提高营销效率的同时，也提高了其风险管理水平。

4) 移动审批

随着移动互联网技术的发展，越来越多的客户选择在网页端口和移动设备客户端口提交授信申请，借助大数据技术的后端审批环节也随之发生了不小的变化。

首先，移动审批实现了客户信息的实时传递。即客户在接入端口填写申请信息时，所填写的申请信息被实时传递给后端的审批系统。其次，移动审批实现了更多的信息采集。基于对大数据技术的应用，客户在申请过程中相关数据也会被系统所采集，如填写时间、修改内容、修改次数、提交时间等信息数据。最后，移动审批的审批过程延伸至申请端。即客户在填写授信申请时，每填写一条信息，该信息就被实时地传递到后台进行核实，客户无须完成全部的申请过程就能得到审批的反馈。

3. 风险预警

1) 风险预警的概念

风险预警是指通过信息的收集和分析，对业务和资产的风险状况进行识别、测量和分析，并对可能发生的风险采取适当措施进行化解，以达到减少损失的目的。商业银行对风险进行预警，可以及时地采取有针对性的措施对未来将会发生的损失进行控制。大数据在风险预警方面极具优势。商业银行借助大数据技术可以从多渠道选取监控指标，对其经营过程中每一个业务的每一个环节的异动进行跟踪，从而实现对风险的有效预警。

风险预警是一个动态过程。在风险预警的动态过程中，主动监测并化解风险是其主要目的，预警是实现该目的的手段。风险预警流程如图 3.9 所示。从图中可以看到，风险预警是一个闭环过程，通过发现问题和解决问题的循环往复实现对风险的动态管理。在这一过程中，监测环节是对风险进行识别的环节，有效的监测识别决定了风险预警的准确性和及时性；预警是触发风险处置措施的环节，而归因分析则是采取恰当处置措施的必要前提；在对当前所发现的风险进行处置后，当即进入下一轮的风险监测环节，以发现新的或变得更加严重的风险问题。

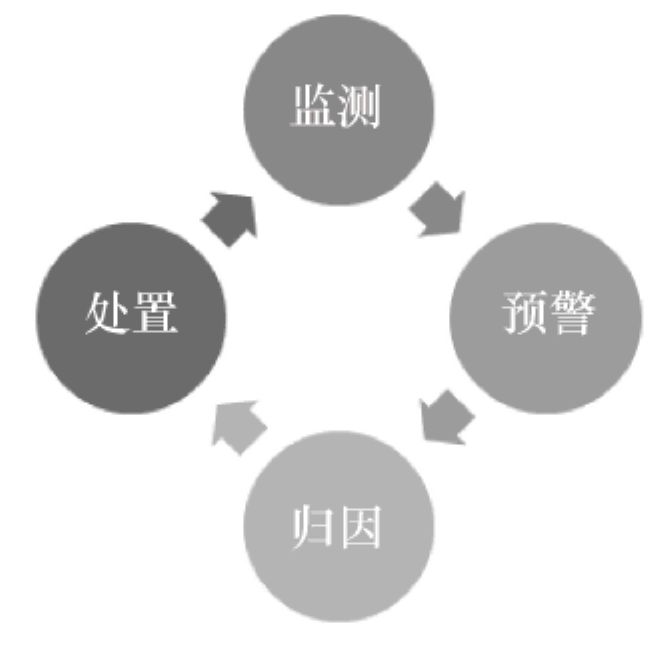


图 3.9 风险预警流程

2) 风险预警体系

健全的风险预警体系是及时且全面的。

风险预警体系的及时性体现在以下两个方面。

(1) 风险预警信号具有前瞻性和预见性。即风险预警信号能够帮助商业银行及时识别早期的风险迹象，避免因预警信号存在滞后性导致其承担较大的损失。

(2) 及时对风险预警信号做出反应。即商业银行在收到风险预警信号后，必须有能力对所发现的风险迹象以化解风险、减少损失为出发点采取快速的应对行动。

风险预警体系的全面性体现在以下两个方面。

(1) 既要关注单一客户，也要关注客户整体。即商业银行对风险预警信号的识别要覆盖到每一个客户个体，也要对整体的客户结构和资产质量给予充分的关注。

(2) 既要细化到单一业务，也要覆盖全部的业务范畴。即商业银行不仅要对微观层面的单一业务进行预警，还要在宏观层面对全部业务的各种风险进行有效的预警和防范。

根据预警类型的不同，可以将风险预警分为个案预警和资产组合预警。个案预警是指对某一客户个体的信用状况的监测和预警；而资产组合预警可以是对某一业务的资产质量的评估和预警，也可以是对由多种业务所组成的整体资产状况的评估和预警。通常情况下，个案预警是资产组合预警的前兆，因此可以在二者之间建立恰当的预警联动机制。

3) 分级预警机制

分级预警机制是指基于预警信号的严重程度和所需响应速度的不同，在预警体系内设置不同的预警级别，以对每个预警信号做出恰当的反应。风险预警信号的分级如图 3.10 所示。不论是哪一级别的预警信号，都需要进行相应的归因分析，在找到预警原因的基础上采取适当的措施对风险进行必要的控制。而分级的意义在于，商业银行可以根据预警信号的级别来确定处置措施的实施范围和实施进度。



图 3.10 风险预警信号的分级

4) 大数据在风险预警中的作用

为提高预警信息的及时性和全面性，商业银行的预警信号获取范围已经扩展到了外部，而且从传统的公共记录扩展到了无限的网络世界当中。互联网大数据具有非常广的数据范围和非常高的数据更新频率，因而基于互联网中快速更新的海量信息的输入，商业银行的预警能力得到了极大的提高。在这一高效运行的风险预警体系下，客户任一异常的行为都会被及时地识别出来，并将作为风险预警信号实时传递给客户经理，客户经理将会根据该预警信号的严重程度采取相应的处置措施，及时对客户的异常情况进行排查。

4. 逾期管理

商业银行是经营风险的企业，因而客户逾期的发生难以避免。因客户逾期所造成的坏账损失是商业银行主要运营成本的一部分；而因客户逾期所收取的逾期利息和相关费用又形成了商业银行的收入。正因如此，对逾期客户进行管理是商业银行风险管理的重要组成

部分。

1) 客户逾期的发生

客户逾期的主要原因可分为：还款意愿差和还款能力不足两个方面。其中还款能力不足是客户发生逾期最为主要的原因，通常可以分为以下 3 种情况。

(1) 客户出现临时性的资金周转困难。若该类客户的还款意愿良好，则其预期的时间不会太长，银行所面临的坏账风险相对较小。

(2) 经济状况恶化导致的还款能力不足。该类客户的坏账风险相对较高。

(3) 贷款金额超过其自身的承受能力。该类客户也同样具有较高的坏账风险。

商业银行利用大数据技术能够从多种渠道获取客户的相关信息，从而能够在事前对存在逾期风险的客户进行有效识别。

2) 客户逾期的处置

对逾期客户的管理主要包括：不良资产处置、逾期催收管理、失联客户管理和逾期信息管理。其中不良资产处置和逾期催收管理是最主要的两个任务。

(1) 不良资产处置。不良资产处置是指通过不良资产核销、不良资产打包出售等方式，对逾期客户所形成的呆账、坏账进行处理，以优化银行资产结构的过程。其中不良资产核销是商业银行处置其不良资产最常见的方式。

(2) 逾期催收管理。逾期催收管理是指商业银行通过采取不同的方式触及客户并实现欠款催回，同时对风险状况不断恶化的客户采取相应的措施，以防范风险敞口的进一步扩大，降低商业银行可能损失的过程。常见的催收方式包括：短信催收、电话催收、实地催收、司法催收等。其中，司法催收的强度最高，所需运营成本也最高；短信催收的强度最弱，所需运营成本也最低。

(3) 失联客户管理。客户失联是逾期客户管理中最常见的问题。在当前的新兴金融模式下，商业银行利用大数据技术，可以对客户的海量数据进行搜集和传递，精准地刻画出客户的个人特征、行为方式和社交网络，进而使其进行真实性核查和风险评估的能力得到大幅提高。在逾期失联客户的管理方面，大数据的作用主要体现为以下两点：一是可以帮助商业银行提前对失联客户进行识别，并在客户失联之前对客户的联系信息进行及时更新；二是可以帮助商业银行利用互联网中所积累的大量关联信息对失联客户的信息进行有效修复。

(4) 逾期信息管理。逾期信息是客户风险预测的数据来源，对逾期数据进行管理有助于商业银行对客户的风险和收益情况做出准确的评价，确定其在客户引入和客户管理方面的具体方向。在逾期管理阶段通过跟踪监测，可以及时地发现客户、流程、授信决策等方面的问题。并且通过对逾期客户管理过程进行检测，可以提高商业银行的运营效率。

3) 逾期催收的计量模型

逾期催收的计量模型是对逾期客户进行分类的重要依据，商业银行可以利用计量模型对客户的风险情况进行识别，进而对不同风险程度的客户采取不同的催收策略和手段。有效的催收策略能够在提高欠款回收率的基础上，降低商业银行的催收成本。常见逾期催收计量模型包括：客户逾期行为模型、账龄滚动率模型和失联模型。

(1) 客户逾期行为模型。

逾期行为模型主要用来对客户未来发生逾期行为的可能性进行预测。由于客户发生逾期行为通常都有一定的表现期，因而客户逾期行为模型通过对客户的交易行为特征和还款行为特征进行分析，考察客户在长期内发生逾期行为的可能性，但对短期(如 1 个月)内的逾期行为预测能力很弱。换句话说，逾期行为模型是对客户资质进行认定的模型，资质较差的客户是逾期客户管理的工作重点。该模型的预测变量主要包括客户的还款行为、消费行为、信用卡取现行为、资金使用情况、欠款情况等方面。

(2) 账龄滚动率模型。

账龄滚动率模型是逾期催收中最常用的计量模型，是在对客户逾期账龄进行确定的基础上，对每一账龄的客户演变到下一账龄的概率进行预测的模型。其中，逾期账龄是指客户未按约定时间还款的违约时间长度，通常以天数来界定，如图 3.11 所示。客户的逾期账龄越高，其违约风险就越大。

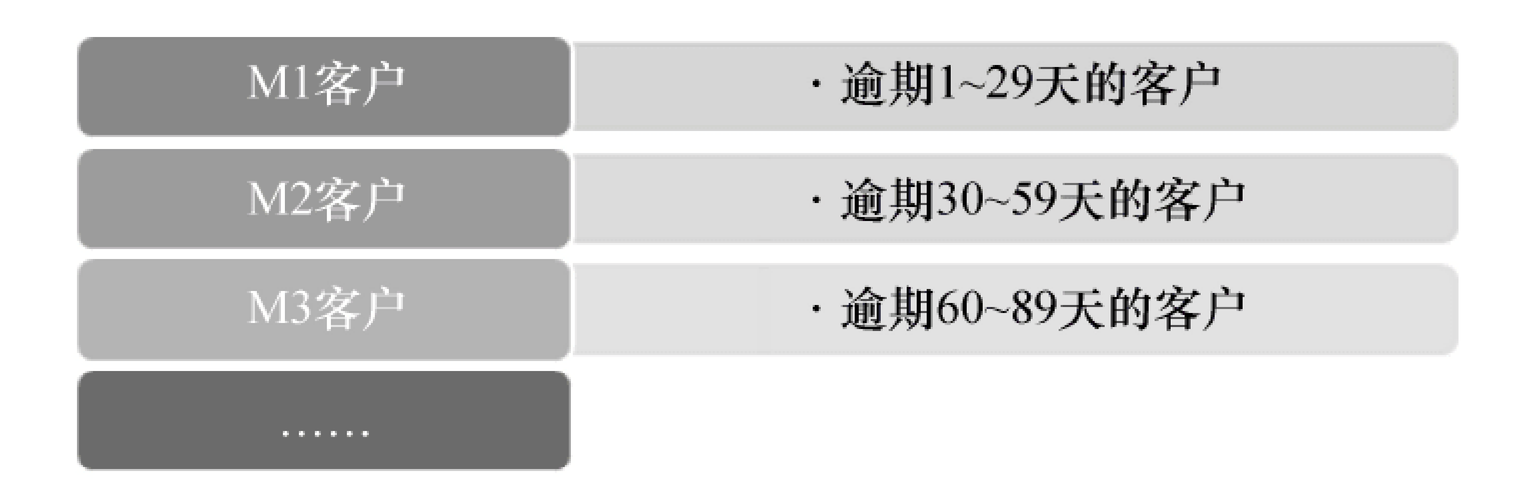


图 3.11 逾期账龄划分示例

该模型体系中通常包含 M0-M1(代表客户从正常客户变为 M1 客户)、M1-M2(代表客户从 M1 客户变为 M2 客户)、M2-M3(代表客户从 M2 客户变为 M3 客户)、M3-M4(代表客户从 M3 客户变为 M4 客户)账龄滚动率模型；通常以逻辑回归模型和决策树等开发方法建立。在该模型中对客户账龄的划分一般都在 90 天以下，因为一旦客户逾期天数在 90 天以上，客户归还欠款的可能性急剧下降，此时商业银行的主要目标是采用严厉的催收方式以尽最大努力挽回损失，无须再考虑制定催收策略。此外，在使用上述模型对客户进行评分时，M0-M1 账龄滚动率模型主要用于对客户风险进行预警和监控；而 M1-M2、M2-M3、M3-M4 账龄滚动率模型则是通过客户评分来区分不同客户所具有的风险，进而对具有不同逾期风险的客户采用不同的催收策略，以最有效的催收方式实现欠款的回收。

在账龄滚动率模型中，预测变量包括行为信息和催收信息两种类型。其中，行为信息包括还款行为、消费行为、信用卡取现行为、额度使用情况等方面；催收信息则包括催收结果、逾期次数、催收后还款行为等方面。在低账龄的客户模型中，行为信息比重较大；而在中高账龄的客户模型中，则为催收信息比重较大。

账龄滚动率模型通常与逾期行为模型结合使用。二者的结合能够帮助商业银行从短期和长期两个方面来识别客户风险，并准确地对客户做出评价，从而做出更具针对性和效率的催收决策。

(3) 失联模型。

失联模型是基于对历史数据的处理和分析，提前预知客户未来发生失联可能性的模

型。客户失联的原因通常有很多，如客户提交虚假材料、恶意贷款、故意断绝与银行的联系、客户信息未及时更新等情形都有可能造成客户失联。因此，商业银行单纯依靠其所掌握的内部信息数据将使模型很难做出正确的决策，需要引入更多更全面的外部数据来提高模型的预测能力。失联模型中的预测变量通常包括客户还款行为、消费行为、贷款余额情况、额度占用情况、联系方式变更情况、历史催收结果、与该客户的联系频率和时间等商业银行的内部信息，也包括客户户籍信息、教育经历、工作单位情况、家庭情况等第三方信息。

失联模型通常与逾期行为模型结合使用。对于失联概率较高且逾期风险较高的客户，商业银行应当给予重视，及时了解客户的实际情况；一旦发现客户失联，及时采取相应措施对客户信息进行修复，将可能遭受的损失降到最低。

4) 催收策略的制定

在大数据背景下，商业银行催收策略的客户细分是基于客户特征的不同进行的。商业银行通过对大量数据进行分析和挖掘，找出不同类型逾期客户的特征，进而有针对性地制定出与客户风险状况相符的催收策略，如图 3.12 所示。

M1客户群	M2客户群	M3客户群	M4~M6客户群	M6+客户群
· 失联：实地催收 · 他行逾期：电话催收风险较高的 · 有意愿无能力：提供展期 · 其他：电话催收、信函催收、短信催收	· 失联：实地催收 · 大额逾期：实地催收 · 有能力无意愿：电话催收风险较高的 · 其他：电话催收	· 失联：实地催收 · 大额逾期：落地催收 · 有能力无意愿：实地催收 · 其他：电话催收风险较高的	· 全部：全面实地催收、司法催收	· 全部：全面实地催收、司法催收、核销

图 3.12 基于客户特征细分的催收策略

在对客户的风险等级进行划分时，通常有以下两种方法：一是根据客户的违约概率对其进行风险评级，即客户无法还款的可能性越高，其风险程度就越高；二是根据客户的风险余额对客户进行风险评级，风险余额是客户无法还款的可能性与客户尚未偿付金额的乘积，即客户的风险余额越大，其风险等级就越高。

在催收策略的制定过程中，需要数据挖掘技术和催收系统的支持。大数据的应用不仅使商业银行对其客户特征的刻画更为精准，而且能够帮助商业银行有效把握不同客户对不同催收手段的反应，从而做出最优的催收决策，提升催收工作的效果。例如，商业银行通过电信运营商的数据发现某逾期客户的电信账户只有在工作日的晚间和周末才有通话或上网记录，那么商业银行在对其进行电话催收时要选在上述时间段内进行。

3.4.3 反欺诈

1. 商业银行所面临的欺诈行为

这里所指商业银行面临的欺诈行为通常是本身带有恶意的目的、到期无还款意愿或虚构偿还能力的借款行为。由信息的不对称所造成的欺诈行为的存在无疑会给商业银行的经

营造成损失。

随着信息技术的不断发展，商业银行贷款业务中所面临的欺诈风险越来越高，主要表现为以下两个方面的原因：一是在当前普遍采用的非现场远程授信模式下，制造虚假申请进而骗取贷款的成本极低。二是我国的信用体系目前还不够完善，客户的违约行为并不会对其日后生活造成重大影响，而以法律手段惩治违约行为的成本又过高，进而导致客户违约的低成本。

根据欺诈行为的发生阶段的不同，可以将欺诈行为分为申请欺诈和交易欺诈两种类型。其中申请欺诈是发生在贷前申请环节中的欺诈行为，包括客户提交虚假的资质证明材料、提交虚假的申请信息和冒充他人身份申请贷款 3 种具体类型。而交易欺诈是发生在贷后阶段中的欺诈行为，包括虚假交易、账户接管和挪用资金 3 种具体类型。

2. 申请欺诈的防范

1) 营销环节

在有营销人员参与的贷款业务申请中，为了防范欺诈行为的出现，营销人员应当在与客户的直接接触中通过与客户进行交流和沟通以及实地走访，对客户的贷款意愿和申请资料的真实性进行核实。对于没有营销人员参与的营销模式下所发生的贷款业务申请(例如通过网页或移动网手机客户端所发起的贷款申请)，对其进行申请欺诈行为的防范主要体现在审批环节。

2) 审批环节

在审批环节中的申请欺诈防范主要体现在以下 3 个方面。

(1) 对客户提交的资质证明资料的真实性进行核实。这么做的目的在于确保客户符合商业银行所要求的授信标准，确保相关信息的真实、完整，方便在贷后环节与客户的联系和互动。

(2) 对客户提交的申请信息的真实性进行核实。这么做的目的在于保证客户的基本申请信息和附属申请信息的真实、可靠。

(3) 对客户借款人身份的真实性进行核实。这么做的目的在于防范申请人冒用他人身份骗取贷款的风险。

商业银行的上述核实过程均属于信息校验，通常包括 3 个层次：一是客户申请信息的逻辑校验；二是客户申请信息与商业银行内部信息间的逻辑校验；三是客户申请信息与外部信息的对比校验。其中利用外部信息进行对比校验离不开大数据的支持，在大数据技术的帮助下，商业银行所获得的外部信息不仅可以用以与客户的申请信息进行交叉比对核实申请信息的真实性，而且还为客户的信息资料提供了有效的补充，并为商业银行提供了更多接触客户的方式。

3) 贷后管理环节

虽然申请欺诈发生于贷款申请环节，但对申请欺诈的防控在贷后管理环节仍需要进行。这是因为依靠贷前的申请和审批环节的审查并不能实现欺诈行为的完全排除。在贷后管理环节对申请欺诈的防控主要有以下两种方式。

(1) 观察客户的违约情况。一些客户的违约行为通常可以反映出其属于申请欺诈，这

些行为特征包括但不限于：早期违约、连续多期未偿还欠款、联系方式在获得贷款后的短时间内即失效等。

(2) 信息的关联排查。通过对客户数据进行关联排查，找出相关数据特征与已发现的申请欺诈数据特征相匹配的客户，对这些客户进行补充调查以核实其申请资料和身份的真实性。

3. 交易欺诈的防范

1) 放款环节

放款环节是防范交易欺诈行为中的账户接管和资金挪用的重要环节。在向客户进行放款时，商业银行通常会对收款账户的户主与该客户的身份进行比对核查，以防止资金被他人使用。此外，商业银行通常还会对客户的贷款账户采取定向支付的资金划转方式，以防范其所贷出的资金被挪为他用。

2) 交易环节

在交易环节防范交易欺诈通常仅存在于商业银行向个人客户所提供的循环授信业务之中，以信用卡产品最为典型。在商业银行所提供的上述个人信贷产品中，客户随时使用额度的过程就是交易的过程。通过对客户的交易过程进行实时或准实时的监控管理，商业银行能够对客户的行为进行及时的观察和有效识别，进而对疑似交易欺诈的行为进行预警。

3) 还款环节

即便客户在当前正常还款，也不能排除其存在交易欺诈的可能。有的借款人为躲避商业银行对其贷款行为的关注和怀疑，会故意正常还款，进而导致商业银行无法对该借款人的真实贷款用途和资质水平做出正确的判断。因此，在贷后还款环节，仍要对客户账户内的资金流向进行监控，进而对交易欺诈行为进行有效防范。

4. 欺诈行为识别模型

1) 申请欺诈模型

申请欺诈模型是通过对客户的相关资料信息进行评分，从而对客户发生欺诈行为的可能性进行判断的。商业银行能够利用其获取的客户申请信息、央行征信信息以及第三方所提供的相关客户信息，对客户的欺诈风险进行评估。

具体来讲，该模型中的预测变量通常包括以下内容：客户的工作单位是否在征信单位列表中、客户的家庭住址和工作地址是否在征信的列表中、申请人是否曾发生过欺诈行为、同一 Cookie 或相近的 IP 地址是否在短时间内多次发出申请请求、发出申请的 Cookie 和 IP 地址是否是该客户经常使用的、客户是否有活跃的互联网行为、客户在电商平台和电信运营商等第三方处所留下的相关有用信息等。

基于大数据的运用，商业银行可以在客户进行线上贷款申请时，就对客户的相关信息进行采集，进而提高模型的准确性。例如，收集客户申请贷款时所处的地理位置，将之与其家庭地址、工作地址进行对比；收集客户在填写申请时的修改内容、修改次数、提交次数等行为信息，将之作为申请欺诈模型的预测变量。此外，大数据还使商业银行的申请欺诈模型和应对策略的信息考察范围得到了扩大，同样有助于提高模型预测和策略制定的准确性。

对于得分低的客户，因其存在很大的欺诈风险，商业银行会直接拒绝该客户的贷款申请；对于得分较低的客户，商业银行则会安排专业审批人员对客户进行二次审查；而对于模型得分较高的客户，因其风险水平较低，商业银行可以通过随机抽样的方法抽取其中部分客户进行风险排查。

2) 交易欺诈模型

交易欺诈模型所衡量的是客户在交易环节发生欺诈行为的可能性。鉴于只有循环授信产品存在交易环节，且信用卡是循环授信产品中最为主要的类型，因而这里的交易欺诈行为，是指不法分子通过盗卡、伪卡等方式盗取客户账户资金的行为(即非法账户接管)以及虚假交易行为。

因为交易欺诈行为具有多样性和隐蔽性的特点，所以交易欺诈模型需要有非常高的精准性。由于神经网络具有很强的自学习能力，能够适应欺诈行为多样且复杂的特点，因而通常商业银行会利用神经网络来开发交易欺诈模型。基于交易欺诈模型的利用，商业银行可以通过客户的历史交易行为刻画客户的行为特征，进而将该客户本次被标识的异常交易与其历史交易行为特征进行对比，若二者间存在较大的差异，则说明发生交易欺诈行为的可能性较大。

该模型的预测变量包括但不限于以下内容：本次交易金额、本次交易时间、本次交易商户、本次交易地点、本次交易币种、过去一定时间内的交易次数、过去一定交易次数内输错密码的次数、过去一定交易次数内交易失败的次数、本次交易 IP 地址、浏览和交易的网站信息等。

由于交易欺诈模型通常要涉及大量的历史交易数据和相关信息并对时效性有较高的要求，因此商业银行利用大数据技术能够很好地对系统的大量运算给予支持。此外，基于对大数据的运用，商业银行能够实现交易欺诈模型中预测变量的及时获取和调整补充，进而使模型的时效性得到有效的保证。

5. 大数据下的反欺诈

1) 用互联网信息描述客户特征

伴随着移动互联网的普及和发展，个人的行为信息越来越多地被记录于互联网之中。商业银行通过利用从多种合法渠道获取的客户在互联网中的相关数据信息(如浏览行为、交易行为、购买记录、搜索记录、社交活动等)，可以对该客户的行为偏好、社交范围、工作状况、文化程度、偿付能力形成一定准确的认知，不再完全依赖于该客户的历史信贷记录和有限的传统审批资料。

2) 线上信息与线下信息相结合

虽然互联网信息可以对客户特征进行描述，但单纯依赖客户的线上信息并不能对该客户形成全面的认知。因此，商业银行只有将内部与外部、线上与线下的多维度信息进行综合使用才能对欺诈行为进行有效的管控。

3) 基于网络技术的非现场审查

贷款申请方式的创新使商业银行的贷款业务越来越多地以非现场的方式开展，但移动互联网技术的发展为商业银行进行远程审查提供了更多的手段。在非现场的贷款业务申请

过程中，商业银行通常会要求申请人通过手机等移动智能设备拍摄包含指定动作的视频或照片，用来对该申请人的身份、工作地点等相关信息进行真实性审查。

3.4.4 反洗钱

反洗钱是政府动用立法、司法力量，调动有关的组织和商业机构对可能的洗钱活动予以识别，对有关款项予以处置，对相关机构和人士予以惩罚，从而达到阻止犯罪活动目的的一项系统工程。洗钱行为会给社会造成诸多不良影响和危害。①洗钱行为会掩盖非法所得、促成资本外逃，进而使贪腐资金转移境外，导致社会财富外流；②不法分子会利用洗钱行为为违法犯罪集团提供资金，因而洗钱行为会助长违法犯罪、破坏社会的和谐稳定；③洗钱行为会动摇社会信用，危害国家金融安全。因此，反洗钱工作在稳定市场经济秩序、阻止非法资金外流、维护社会稳定中发挥着重要作用。

商业银行是反洗钱职责的主要承担者。在全球经济一体化和信息化不断加快的背景下，洗钱犯罪的特征也呈现出隐蔽、快速的新特点。在大数据时代，随着大数据技术的日趋成熟和完善，商业银行也开始将大数据技术应用到防范和控制洗钱活动、提升反洗钱工作的效率中来，通过构建统一的反洗钱工作系统，对商业银行所拥有的内部海量数据进行充分整合和深入挖掘，进而使反洗钱工作的时效性和准确性得到提高。

1. 大数据在反洗钱工作中的优势

1) 发挥商业银行的数据优势

在商业银行开展业务的过程中，每天都会产生海量数据。这些数据包括商业银行交易系统中所产生的海量交易信息、商业银行业务处理流程中用于作业和授权的影像资料等半结构化数据以及客户的投诉和评价等交互信息。因此，商业银行在对大数据进行应用方面具备天然的优势。商业银行通过充分利用大数据技术与聚类、神经网络、决策树等智能算法，能够对其所掌握的数据进行有效的分析和挖掘，进而提升其自身在反洗钱工作中的时效性和准确度。

2) 提高反洗钱调查的时效性

商业银行在进行反洗钱调查时，主要依据《金融机构大额交易和可疑交易报告管理办法》对客户身份的真实性进行识别。只要相关交易的数据特征符合可疑交易的给定标准，商业银行就会将该交易数据报送至反洗钱监管机构。商业银行在判别客户交易是否具有可疑性时，只有在客户身份真实性识别的准确度得到提高的前提下，才能实现其可疑性审查质量的提高。在数据的应用下，商业银行在对客户身份的真实性进行审查的过程中，可以将可疑交易数据与客户所在地域、工作状况、受教育程度、收入水平等个人身份特征相联系，进而减少可疑性审查出现失真和误报的可能性，提高反洗钱调查的实效性。

3) 提升反洗钱工作的效率

商业银行内部有许多信息系统，这些信息系统是分散且异构的，各个信息系统的技术指标也不尽相同，因而导致每个信息系统都是封闭的信息孤岛。正因如此，基于上述关系型数据库和传统数据挖掘技术所构建出的反洗钱工作系统，会面临大量数据的格式不统一、无法存储、难以处理等技术障碍。由于大数据技术能够对非结构化数据进行处理并允

许数据存在不一致，因而利用大数据技术可以解决上述传统反洗钱工作系统中所存在的难题，缩短系统的响应时间，进而使商业银行在反洗钱工作中的效率得到提升。

2. 商业银行基于大数据的反洗钱工作系统

1) 反洗钱工作系统的工作目标

反洗钱工作系统的工作目标主要包括以下 4 个方面：一是构建基于大数据的数据仓库；二是对数据进行加载、处理、清洗、转换；三是配置反洗钱业务规则；四是对可疑数据进行展示。

2) 反洗钱工作系统的逻辑分层

(1) 源数据：商业银行内部各个系统中的数据。

(2) 数据存储：在初始状态下与源数据层的表结构一致，但之后不再随原数据层表结构的变化而变化。

(3) 数据汇聚：完成对客户、账户和交易数据中的相关主题数据的采集和整理。

(4) 数据分析：根据预先设定的可疑规则对数据汇聚层的数据进行计算分析，从中找出可疑交易并生成可疑报表。

(5) 信息管理：对数据分析层所得出的预警信息和报表信息进行管理。其中具体包括：用户管理、规则定义、权限管理、日志管理、报表管理、报送管理等相关管理活动。

(6) 决策分析：商业银行相关工作人员对预警信息进行处理以对可疑交易进行确认，进而将所筛选出的可疑数据报送相关监管部门。

3) 反洗钱工作系统的系统架构

反洗钱工作系统的系统架构如图 3.13 所示。

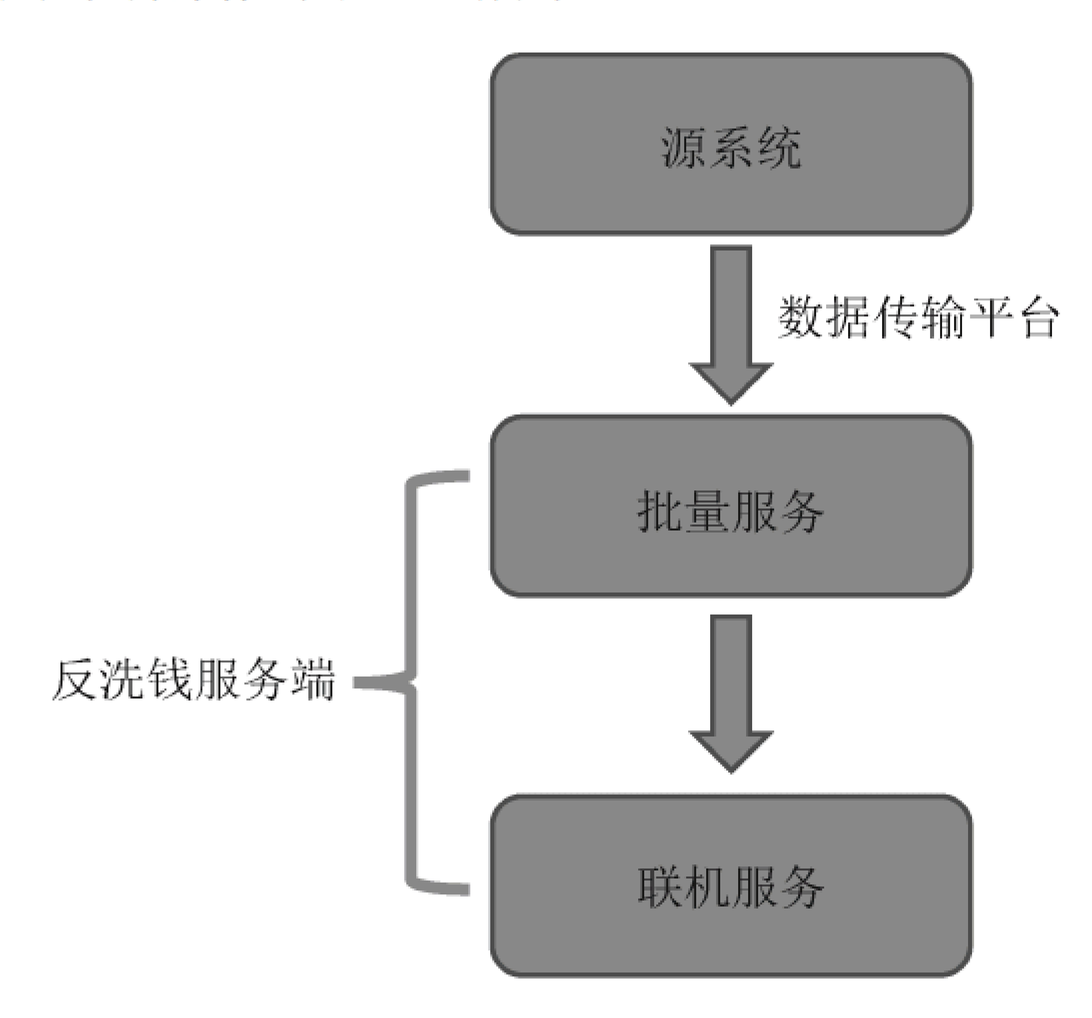


图 3.13 反洗钱工作系统的系统架构

(1) 源系统：即属于源数据层的商业银行内部的各个数据系统。

(2) 数据传输平台：该平台的功能在于将源系统中的数据传输至指定位置。

(3) 反洗钱服务端：服务端由批量服务和联机服务两部分组成。其中，批量服务是指系统自动完成对可疑数据提取的活动，即从源系统获取源数据，并进行清洗和装载；而联

机服务则是指系统用户进入系统完成系统操作的活动。具体来讲，批量服务在源数据的基础上根据预设的各项指标计算账户和客户的指标值，在各项指标值的基础上根据预设的可疑规则生成可疑报送数据，同时进行报表生产，属于系统逻辑分层中的数据存储层、数据汇聚层、数据计算层。而联机服务则主要包括系统管理、参数设定、预警 / 报告处理、统计报表，属于系统逻辑分层中的信息管理层。

3.5 运营优化

随着中国经济进入“新常态”以及互联网金融的快速发展，商业银行在当前面临着巨大的冲击和挑战。从发达国家银行业的发展经验来看，利用大数据挖掘能够大幅提高商业银行的经营水平。

从运营角度来看，大数据技术为商业银行透彻地了解市场创造了可能。商业银行可以通过对海量客户行为数据进行分析，了解未来市场的发展方向，进而做出有针对性的战略安排。

3.5.1 市场和渠道分析优化

1. 商业银行的市场渠道

1) 商业银行市场渠道的种类

商业银行有柜台网点和电子银行两种市场渠道。其中，电子银行是指商业银行等银行业金融机构利用面向社会公众开放的通信通道或开放型公众网络，以及银行为特定自助服务设施或客户建立的专用网络，向客户提供的银行服务。商业银行所开发的自助银行、网上银行、电话银行、手机银行、微信银行等都属于商业银行的电子渠道。

柜台渠道是商业银行最传统的市场渠道，能够使商业银行工作人员与客户面对面地进行接触和沟通，业务范围也最为广泛。截至 2015 年年末，我国商业银行的业务网点已有 22 万家，并分布在世界各地。

自助银行主要表现为 ATM 集群。ATM(Automatic Teller Machine)也称自动柜员机，是在无人直接干预管理的情况下，能够向客户提供查询、存取款、转账汇款等金融服务的银行自助设备。作为对柜台渠道的补充，自助银行使商业银行的经营效率得到提高、业务覆盖面得到扩大。

网上银行是指商业银行依托互联网向客户提供相关金融服务的方式。客户能够随时随地访问商业银行的网页，并根据网页上的导航和操作指引办理其所需要的业务。

电话银行是指商业银行以语音通信的方式向客户提供相关金融服务的方式。客户可以随时随地通过商业银行所提供的自助语音服务和人工服务办理其所需要的日常业务。

手机银行是指商业银行以安装在客户手机内的客户端软件为媒介，为客户提供相应金融服务的方式。手机银行与上述商业银行的市场渠道相比具有更高的便捷性。

微信银行、微博银行等创新渠道是商业银行以社交软件为媒介为客户提供相应的金融服务的创新方式。其中，微信银行对商业银行的市场拓展作用最为明显。在这一创新的市

场渠道中，客户不仅可以办理所有的日常业务，还能够及时且直观地接收到商业银行的活动推广。

2) 商业银行市场渠道的新特点

在大数据的背景下，商业银行的市场渠道在发展中也呈现出新特点。

(1) 运营模式互联网化。随着移动互联网的快速发展，越来越多的人活跃在网络世界中。作为以客户为服务导向的商业银行也开始重视对电子渠道的发展，尤其是对手机银行、直销银行等创新渠道的利用。

(2) 客户定位精准化。由于商业银行通过不同的市场渠道所接触到的客户群存在一定的不同，且越来越多的客户信息能够被采集和利用，因此商业银行借助大数据技术在对不同的市场渠道进行利用时，对各个渠道客户特征的描述也日益精准，对市场的敏感性也不断增强。

(3) 市场战略差异化。在对各个渠道的客户群做出精准定位后，商业银行针对各渠道客户群的不同特征制定出不同的市场战略。客户定位越精准，市场战略的差异化越大。

2. 市场渠道的运营质量

对不同市场渠道的运营质量进行监控，可以帮助商业银行筛选出发展前景较好的渠道，以对该渠道进行进一步的调整和优化，进而增强商业银行的市场竞争力；还可以帮助商业银行分析出各个渠道适宜推广的产品和服务，进而实现产品和服务推广策略的优化。

在大数据的背景下，对渠道运营质量进行考量已不再是一件难事。对渠道运营质量的考量可以从成本、收益\客户的感知和偏好 3 个方面进行。

1) 成本方面

市场渠道的运营成本通常由商业银行在该渠道开发和使用的过程中所投入的全部价值量构成。一般包括商业银行所投入的人力成本、营销成本、IT 运营成本和开发成本。其中，IT 运营成本包括硬件成本、软件成本、维护成本和 IT 设备折旧等成本；开发成本则包括产品的设计和开发、业务开发、软件开发、系统开发等相关活动的成本。电子渠道与柜台渠道相比，在降低成本方面的表现更为突出。

2) 收益方面

市场渠道的收益通常表现为商业银行利用各渠道提供金融服务所实现的收入，例如转账收益等相关收入。商业银行创新的市场渠道(如微信银行)与其传统市场渠道相比，还具有一定的获客优势，可以在为存量客户提供服务的同时吸引更多的新用户加入，进而为商业银行带来更多的收益。

3) 客户的感知和偏好

由于客户的反馈是直接的市场反应，因此在对市场渠道的运营质量进行考量时，还需要考虑客户在各渠道中的感知和偏好。商业银行可以通过分析客户主动做出的评价反馈以及客户在该渠道中发生的行为获取客户的感知和偏好。

3. 市场渠道运营优化

大数据在商业银行市场渠道运营优化中的作用主要表现为以下 5 个方面。

1) 构建 360 度全景客户信息视图

大数据具有强大的信息获取、集成和分析的能力。因此，商业银行将其内部的客户相关数据与外部第三方所拥有的信息数据相结合，利用大数据技术可以对客户从多个角度进行认知，进而构建起 360 度全景客户信息视图，以帮助商业银行根据不同客户的不同需求对其市场渠道进行有效优化。

2) 实现精准化营销和精细化服务

在大数据的背景下，商业银行可以根据客户的身份背景、行为特征等多方面信息对每一位客户的服务需求进行预测和判断，进而向每一位客户进行有针对性的营销、提供有针对性的服务。在这一过程中，商业银行市场渠道的运营效率得到了提升。

3) 增强客户黏性

商业银行在利用大数据技术对客户形成准确认知的基础上，可以依据客户特征的不同对客户进行精准营销，进而使客户关系得以深化，使客户的渠道交易活跃度以及产品覆盖度得以提高。

4) 提升客户对商业银行的价值贡献

在客户黏性增强的基础上，客户无疑会为商业银行带来更多的价值贡献，进而使渠道收益得以增长，使渠道质量得到有效的提升。

5) 发现新的市场需求

基于大数据的利用，商业银行可以及时地捕捉到潜在的市场需求，进而可以根据该需求对其全部或部分市场渠道的业务功能进行补充和完善。

3.5.2 产品和服务优化

1. 产品策略的优化

1) 产品研发个性化

在大数据背景下的产品研发过程中，商业银行可以将客户行为转化为信息流，在对各类型客户的身份背景、行为偏好等进行深入了解的基础上合理预测客户需求，进而根据上述大数据分析的结果有针对性地为不同的客户群制定不同的金融产品，从而为客户提供最优的产品体验。

2) 产品设计模块化

产品设计的模块化是指商业银行根据大数据分析的结果，设计出多种不同的定制化模块并构建出模块化的产品选择体系。在产品设计模块化的基础上，商业银行可以根据不同的客户需求和营销场景对其产品和服务进行个性化搭配，从而为客户提供最佳的金融产品与服务的组合。

2. 价格策略的优化

1) 价格策略的差异化

基于对大数据技术的利用，商业银行可以在结合自身市场定位的基础上，根据客户层次和需求的不同为客户提供差异化的价格策略。即商业银行可以通过利用大数据技术整合和分析海量数据对市场和客户进行有效细分，进而根据不同的客户类型、不同的业务类

型、不同的行业、不同地区的风险程度确定不同的业务收费标准。

2) 价格策略的动态化调整

由于市场和客户需求都是在不断变化的，因此商业银行需要根据这些变化来调整其价格策略以助其获得有利的市场竞争地位。商业银行利用大数据技术可以对金融市场的资金面状况以及其他投资品价格的实时数据进行分析，进而对未来利率走势做出预测，以便对其产品和服务的价格做出及时的调整。此外，基于对大数据技术的应用，商业银行可以深入了解不同客户的需求及价格敏感度，进而对其价格策略进行有针对性的调整，在赢得新用户的同时对存量客户进行有效维护。

3. 客户服务的效率优化

大数据背景下、商业银行客户服务的效率优化主要体现在商业银行向其客户提供的个性化增值服务上。其中，个性化增值服务主要包括个性化的产品推荐、位置营销、电子渠道的全景体验等服务。

商业银行通过利用大数据技术所获取的信息数据，可以了解到客户密切关注或频繁访问的特定事件，进而结合通过分析客户购买行为所得出的全方位需求预测，及时地向客户提供相应的服务推荐和优惠信息。在这一增值服务提供的过程中充分地实现了对客户的尊重，能够有效地获得客户基于其价值自我实现的认同。从中可以看出，基于大数据技术的客户服务与传统方式相比，服务效率得到了明显的提高。

3.5.3 网络舆情分析

1. 商业银行网络舆情的类型

1) 根据发展过程划分

根据发展过程的不同，商业银行网络舆情可以分为渐进式和突发式两种类型。

(1) 渐进式网络舆情。是指发展过程较慢、矛盾在网络中逐渐积累并最终由某一事件触发的舆情。例如，客户在商业银行网点办理业务排队时间过长所导致的舆情事件即为渐进式网络舆情。

(2) 突发式网络舆情。是指发生得十分突然，且在网络中快速传播并引起公众的强烈反应的舆情。例如，商业银行的交易系统突然发生故障使客户资金无法及时到账所引发的舆论事件即为突发式网络舆情。

2) 根据成因不同划分

根据发生成因的不同，商业银行网络舆情可以分为诽谤型、误解型和情绪型。

(1) 诽谤型网络舆情。是指不法分子为谋取不正当利益所进行的恶意造谣和诽谤。例如，在网络上发布不实消息称某商业银行 ATM 吐假钞所引发的舆情事件。

(2) 误解型网络舆情。是指客户基于其对商业银行相关规章制度、业务行为的重大误解，在网络上发布言论抨击商业银行所引发的舆情事件。例如，有小额取款需求的客户认为银行工作人员建议其去 ATM 上取款是歧视行为。

(3) 情绪型网络舆情。是指客户基于对商业银行在为其提供金融服务时存在的疏忽和纰漏的不满，在网络上发布相关言论所引发的舆情事件。例如，客户因在柜台办理业务时

所耗费的时间过长所产生的不满情绪。

2. 商业银行网络舆情的监控系统

商业银行可以利用大数据技术建立起网络舆情监控系统，进而帮助商业银行提高其网络舆情的风险管理能力。网络舆情监控系统可以自动搜集和分析潜在的舆情信息，及时发现存在的风险因素，进而对风险进行有效的预警。此外，网络舆情监控系统还可以帮助商业银行对已发生网络舆情事件的发展态势进行监测，并对舆情控制措施的实施效果进行检验。因此，商业银行的舆情监控系统通常包括以下 4 个模块。

1) 网络舆情的信息收集模块

通过利用大数据技术在网页、论坛、社交平台等网络媒介中根据事先设定关键词对网络舆情进行收集和整理，并将所获取的网络舆情传送至信息处理服务器中。

2) 网络舆情的信息集成和分析模块

在该模块中，商业银行所获取的网络舆情信息将被进行再次筛选和甄别，进而实现数据噪声的有效剔除；此外还将根据发生频率的不同对筛选后的舆情信息进行分类和初步分析。

3) 网络舆情的风险评估模块

根据对各网络舆情所表达观点的倾向性进行分析和统计，进而得出各个舆情的公众关注度和风险程度，并对舆情的发展趋势做出初步的判断。

4) 网络舆情的风险报告和预警模块

该模块将会对网络舆情风险评估模块所得出的结论进行进一步的分析和总结，从舆情的性质、危害程度、影响范围、可控程度等角度对该舆情所具有的风险进行量化，并根据量化结果对超出预警值的舆情进行预警。如图 3.14 所示为网络舆情监控系统。

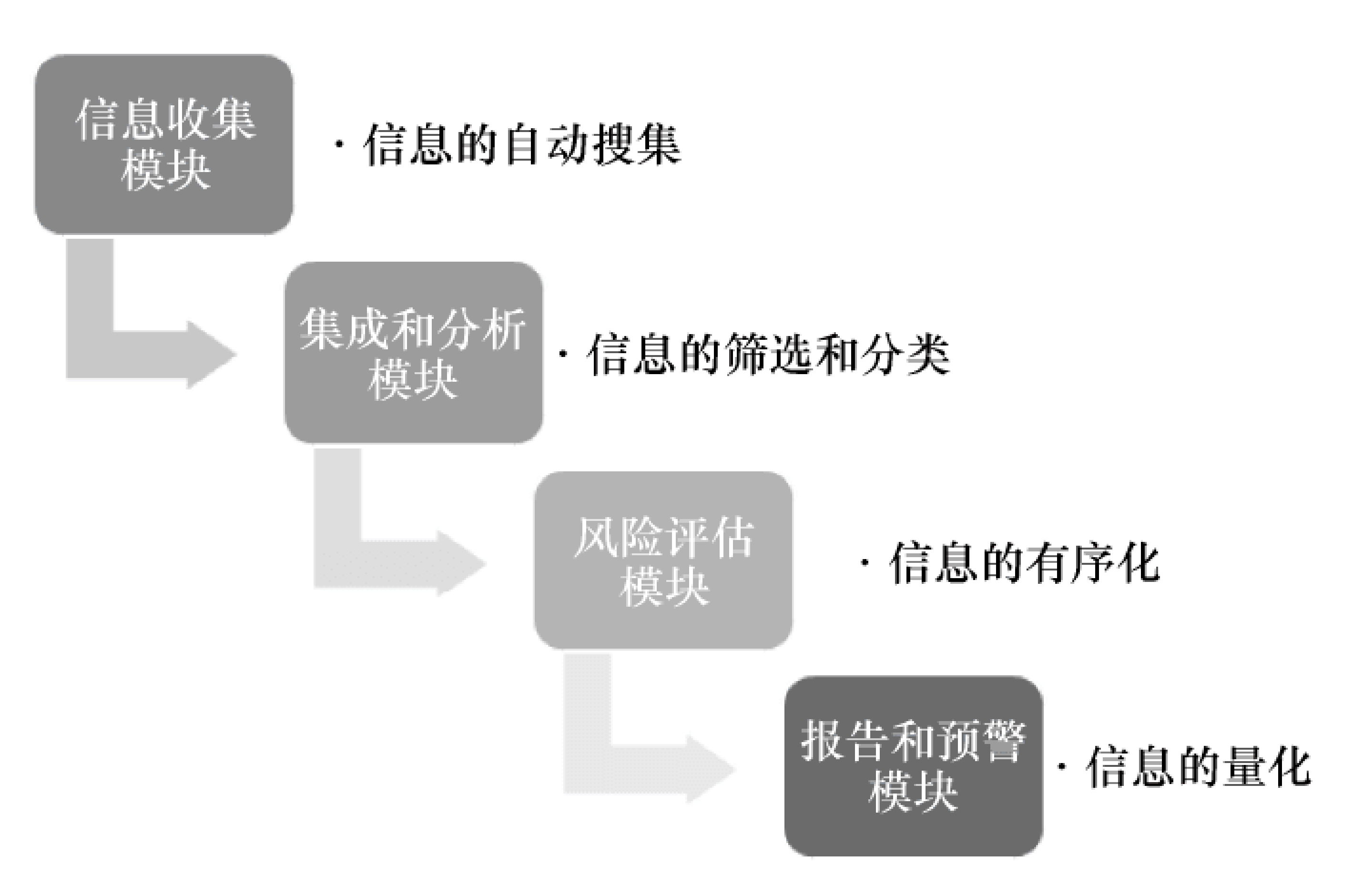


图 3.14 网络舆情监控系统

3. 大数据技术在商业银行网络舆情分析中的作用

1) 客户洞察

商业银行在利用大数据技术对网络舆情进行分析时，可以基于其多方获取的网络信息

对客户的特征、需求、搜索关键词等多种客户网络行为特征进行全天候的实时监测，从而实现准确的客户洞察。在准确洞察客户的基础上，商业银行能够进行有效的客户关系管理并做出准确的营销决策。

2) 舆论导向

基于对大数据技术的应用，商业银行在对网络舆情进行监控时能够及时地获取该话题的舆论导向，使其对网络舆情的把握和控制能力得到有效的提高。对于正面的网络舆情，商业银行可以基于对舆论导向的把握在合适的时机对自身进行恰当的宣传；对于负面的网络舆情，商业银行可以基于其对舆论导向的控制能力适时地采取应对策略，尽最大努力阻止事件的进一步发酵，以将负面影响降到最低。

3) 竞争动向

随着我国金融领域改革的不断深化，金融业的竞争也日益激烈。因此，竞争对手的一举一动都有可能会牵动商业银行的运营安排、市场战略的实施和调整。商业银行利用大数据技术对网络舆情进行监控，不仅可以及时获取自身的相关信息动向，也可以及时获取竞争对手的信息动向。对竞争对手的网络舆情进行分析可以为商业银行制定产品战略、优化客户服务等活动提供客观的参考依据。

3.5.4 案例——大数据分析助力手机银行优化创新

随着手机银行产品的不断同质化，拼功能、拼免费的阶段已经逐步退去，质量过硬的产品，极致、极简的客户体验，是接下来差异化发展的趋势。通过大数据“深度学习”制定追踪指标，可以辅助产品经理更好地挖掘用户需求、洞悉客户喜好、评估产品性能，实现产品创新优化。

江苏银行新版手机银行上线以来，极力打造“只为你、更懂你的手机银行”，非常重视相关运营数据的收集和分析工作，实时监控后台运行，保证产品性能稳定，动态掌握用户行为，分析功能受众程度，从数据中挖掘用户体验反馈，评判产品功能是否真正戳到了客户的痛点，为手机银行的优化和创新提供了强有力的驱动与支撑。

1. 性能跟踪

对于任何一项产品，性能稳定是基础，没有稳定的性能保障，多样的功能服务、流畅的 UI 交互皆是空谈，相当于直接把客户拒之门外。对此江苏银行监测以下数据，及时发现并解决客户使用中出现的问题。

(1) 安装、启动异常。监测安装、启动手机银行 APP 过程中存在的问题，记录报错机型、系统版本、APP 版本，便于分析解决问题，诸如手机银行未能及时更新以适用最新版本的手机系统，不支持低版本手机系统或者对手机剩余存储空间有要求。

(2) 加载时间。监测页面加载时间，尤其是首屏。据 TalkingData 统计分析，使用手机时，如果页面加载超过 5 秒钟，74%的客户会选择离开，因此，应尽量控制首屏加载在 450K/s 内。若部分页面加载速度明显过慢，需分析原因并加以改进；若因客户网速导致加载缓慢，应及时以亲和的语言予以提示，客户一般很难区分、也不愿主动去思考其无法顺畅使用的原因，只会因此对产品失望而选择离开。

(3) 使用中的异常、闪退。监测运行中异常、闪退发生的页面及次数，记录报错机型、系统版本、APP 版本、提示信息，便于分析解决问题。

(4) 日登录客户数。每日登录过手机银行的客户数(区分新登客户、存量客户)。掌握手机银行日登录客户数的动态变化，监控低谷，原因是否与版本更新、系统故障等有关。

(5) 使用时长。统计客户单次操作手机银行的时间，分析会话时间低于某一阈值的次数占比，结合时间点以及其他数据分析可能原因。

2. 业务流程优化

客户使用手机银行产品的基本诉求就是顺畅完成目标交易，如果预先没有相关业务提示或者提示不明显，导致客户操作过程中、甚至提交之后才发现不能继续交易，就会产生浪费其时间、感情的负面情绪；另外，如果业务操作流程不够人性化，用户交互步骤不够精简，也会给客户带来急躁甚至愤怒的心理反应，对此江苏银行做了以下数据追踪。

(1) 业务中断提示。监测业务不能继续的提示页面及次数，便于优化流程，预先做好客户引导，提升体验。

(2) 漏斗分析。记录每项菜单功能每个步骤页面的访问次数，停留时长，计算每步的转化率，根据最后停留页面，找出客户未办结业务或快速离开(停留时间低于规定阈值)的原因。

(3) 跳出率。监测客户未办结即退出手机银行的菜单页面和次数占比，有可能该项功能的体验存在较大问题导致客户弃用。

(4) 关联操作。记录客户成功办结一项业务之后的页面操作，记录客户离开当前菜单后进入的下一个页面，为产品设计的进一步优化提供思路。

3. 功能优化创新

所有的产品功能都应以客户为中心，如果不被客户接受并所需，即为零。对此江苏银行做了以下数据统计。

(1) 菜单点击量、平均停留时长。掌握客户常用功能，挖掘客户潜在需求，优化现有产品，加强相应功能化或场景化产品拓展。

(2) 交易次数、人数。记录每项业务交易次数、人数(区分办结/未办结)，尤其关注版本更新、新功能上线、功能优化、营销活动发布等时间节点的交易次数、人数的动态变化，对比同时间段的菜单点击量，若点击量远大于交易次数，分析客户对新功能或活动接受度不够理想的原因。

(3) 登录时间。计算各个时间段使用手机银行客户占比，掌握峰期、谷期，以便调整信息、活动的推送时间，进一步对登录时间相对固定的客户，实现更适时的推送。

(4) 推送信息的曝光次数或阅读量。记录新产品或营销活动上线，告知客户时客户的阅读量和引导提示曝光次数，结合产品同时间段的点击量、交易量变化，对比分析产品或活动的宣传效果。

(5) 分享次数。对于具有分享功能的页面，记录客户分享次数；记录截屏页面及相应

次数。

未来的银行不只是数据的银行，更要是数据分析和数据解决的银行。做一款有温度的、让客户留恋的手机银行，同样也离不开数字化运营。无论是领先的产品创新、到位的功能优化还是精准的营销管理、实时的风险监控，都需要从数据中摄取价值。本文简述的运营数据分析仅走出了“大数据”学习的一小步，加强内外部数据整合能力、数据分析应用能力和数据辅助决策能力将成为量质并举的前提下争夺移动端市场的关键。江苏银行手机银行秉承以客户需求和体验为中心，积极拥抱“大数据”技术，不断超越、突破，为客户打造了一个即需即用、便捷贴心的移动银行。

本章总结

- 商业银行在长期的金融服务中，积累了大量的信息数据，这些数据涵盖了客户的个人基本资料、收入情况、生活方式以及过往接受金融服务的历史记录等相关资料；可以通过利用先进的数据库系统和大数据挖掘及分析技术，对其所掌握的客户信息进行充分的利用，进而实现多个维度的客户细分。
- 在大数据技术的应用下，商业银行可以及时发现客户尚未被满足的需要和对现有服务的不满，及时采取恰当的行动解决客户的诉求，从而在客户结束其与银行的业务关系之前，及时对客户进行挽留，最大限度地减少客户的流失。
- 在商业银行向客户提供增值服务的过程中，通过应用大数据技术能够发现客户尚未被满足的服务需求，从而有意识地完善和提高客户在商业银行各渠道中的服务体验，提高客户黏性。
- 大数据在商业银行的客户生命周期管理中充分的应用，进而能够帮助商业银行进行实时营销、交叉营销、社交化营销和个性化推荐。
- 大数据可以帮助商业银行在创新模式下进行贷款风险评估。通过利用大数据技术商业银行能够从多个维度获取客户信息，并利用有效的风险计量技术对其所面临的贷款风险进行合理评估。
- 在大数据技术的应用下，商业银行可以对信用卡客户的信贷风险进行实时监控，并根据监控结果及时对客户的授信额度做出调整。
- 大数据风险控制的优势主要体现在大数据征信的利用价值上。第一，大数据使商业银行的客户信用风险评估纳入了多样化的行为数据，这些数据覆盖范围广泛且具有实时性；第二，在大数据风险控制中，信用评价更加精准；第三，大数据风险控制中对客户信用的评判更具时效性。
- 大数据在商业银行的反欺诈和反洗钱工作中也能发挥其在数据处理和分析中的独特优势。
- 大数据能够很好地帮助商业银行对其市场和渠道分析、产品及服务进行优化，还能够帮助商业银行对网络舆情进行分析从而优化其日常运营。

本章作业

1. 客户细分都有哪些类型？

2. 大数据是如何帮助商业银行进行客户流失预测、渠道管理优化和提供增值服务的？

3. 什么是客户生命周期管理？大数据在其中如何发挥作用？

4. 简要概括大数据如何帮助商业银行进行实时营销、交叉营销、社交化营销和个性化推荐。

5. 传统的贷款风险评估面临哪些挑战？大数据又是如何帮助商业银行进行贷款风险评估的？

6. 简述大数据在信用卡自动授信中是如何应用的。

7. 大数据风险控制与传统风险控制有哪些区别？大数据是如何帮助商业银行进行风险管理的？

8. 简述大数据是如何在商业银行反欺诈和反洗钱工作中发挥作用的。

9. 大数据如何实现对商业银行运营的优化？

第 4 章 大数据在证券行业中的应用

本章目标

- 掌握大数据技术在股价预测中具体应用
- 熟悉如何用大数据技术进行证券客户关系管理
- 了解大数据技术在投资情绪分析中的应用
- 掌握大数据技术在证券量化投资方面的应用

本章简介

在金融行业中，证券业属于数据密集型行业，积累了上市公司财务报表、客户关系、市场信息、交易数据等大量信息，伴随着时间的增长和上市公司数量的不断增加，其数据已呈指数型增长趋势。而这些数据的分析和处理对投资者、券商乃至整个证券市场来说是至关重要的。例如，一家券商发布的一份股票研究报告就很可能会影响投资者或者其他券商的投资决策，进而对整个证券市场产生影响。券商为了应对激烈的同业竞争，也都争相把大数据技术作为保护自己市场地位的有力武器。随着大数据技术的成熟和证券市场的网络化，大数据目前已经应用于证券行业的方方面面。本章主要从股票分析、客户关系管理、投资情绪以及量化投资方面出发，介绍大数据技术在证券行业中的应用。



4.1 大数据在股票分析中的应用

股票分析主要分为技术分析和基本面分析两大类，其中技术分析主要由交易策略和买卖时机构成；基本面分析主要由股票选择和投资组合构成。大数据技术的应用主要体现在数据挖掘上，在基本面分析方面，主要运用的是决策树、聚类分析两类研究方法；在技术分析方面，主要运用的是人工神经网络(BP)、基因遗传、决策树、关联分析等。

在数据分析时，一般会以某段时间中国宏观股市数据或上市公司相关资料为基础，运用 SPSS、SAS 等工具对数据进行处理、算法改进以及数据挖掘工作。根据所采用的挖掘方法的不同，所处理方法如下。

4.1.1 基于基本面分析的数据挖掘方法

基本面分析，广义上是指以经济学的供求关系原理为基础，通过以判断金融市场的未来走势为目标对历史的经济和政治数据进行分析。分析因素主要有宏观经济状况、利率水平、通货膨胀、企业素质、政治因素等。狭义的基本面分析通常是指广义基本面分析中的企业素质，分析因素主要包括企业财务报表、行业状况、管理层素质、产品的市场竞争力等，如表 4.1 所示。

表 4.1 基本面分析中的主要分析因素

基本面分析	主要因素
企业财务报表分析	市盈率/市净率/净资产收益率/流动比率/销售净利润； 每股收益/每股净资产/每股利息分配； 成本费用率/负债比率。现金比率/应收账款周转率
行业分析	行业类别/行业成长度
企业产品市场竞争力分析	市场占有率/市场价格/销售能力/原材料价格
公司文化和管理层素质	管理层能力/企业内部协调能力

1. 决策树

以 ID3 算法为主，按照投资者所感兴趣的指标(财务比率、流通比例等)来挖掘出符合投资者需求的上市公司。为了寻找其规则和先后顺序(建立决策树)，首先将资料按照投资者需求做数据预处理；然后预处理后的资料分成训练样本和测试样本，训练资料中以 Gain 最大者为根节点，建立决策树，再由决策树建立分类规则；最后以分类规则寻找股市中符合要求的公司。

2. 关联分析

以划分算法为主，对于投资者所感兴趣的股票，根据各项财务指标对其进行分析，从中找出最佳的投资组合。

3. 聚类分析

以自组织映射(SOM)聚类算法为主，对于投资者给定的一组具备属性值的个股资料，找出一个能够按照属性值将个股聚类的模式，使得属于同一聚类内的个股的相似性最大

化，不同聚类间的个股相似性最小化，并分析出没在个股属性中显示而是隐含在各聚类中的共同特性。

4. 人工神经网络

以投资者感兴趣的财务指标、负债情况、盈利能力为分析变量建立前向式的神经网络模型，并通过分析找到最佳投资组合。

5. 逻辑回归

以投资者提供的个股基本面指标为变量建立二元逻辑回归模型，从个股中找出最佳投资组合。

4.1.2　基于技术分析的数据挖掘方法

1. 决策树

以 C4.5 算法为基础，首先将投资者对股票买点的规则要求作为分类样本，将买点分类为“+”、“-”两类群体，再将投资者所需要分析的指标作为自变量，最后利用决策树产生的“+”类群体的分类规则来找出自变量的有效区间并从中筛选出“有效买点”。

2. 人工神经网络

以 BP 算法为主，由投资者提供的个股历史价格数据集通过训练—学习的循环过程来预测未来某一时间段的价格，提示投资者最佳入场时机。

3. 时间序列分析

按照投资者指定的个股和板块指数，对其价格走势进行分析并建立 ARIMA 模型，利用历史价格变动来预测在未来一段时间内的价格走势。

4. 关联分析

以 Apriority 算法为基础，通过投资者给定技术指标以及投资者指定的个股历史信息，找到能够以其中某些指标的出现与否来预测其他指标出现与否的规则。各类算法在分析与预测中的作用如表 4.2 所示。

表 4.2　基于基本面分析和技术分析的算法分类

股票分析与预测 / 数据挖掘算法	基本面分析		技术分析	
	股票选择	建构投资组合	选择投资策略	选择买卖时机
决策树	○	×	○	○
人工神经网络	○	○	○	×
遗传算法	×	○	○	×
聚类	○	×	×	×
逻辑回归	×	○	○	○
时序模式分析	×	×	×	○
关联分析	×	○	○	○

注意：○代表可行，×代表不可行。

接下来介绍主要的 3 种分析方法：决策树法、聚类分析法、人工神经网络算法。

4.1.3 决策树法的应用

决策树算法(Decision tree)是一种逼近离散函数值的方法：它是一种典型的分类方法，首先对数据进行处理利用归纳算法生成可读的规则和决策树，然后使用决策树对新数据进行分析。本质上决策树是通过一系列规则对数据进行分类的过程。它的树状结构由根节点、内部节点、分支以及叶节点构成。整棵树的结构分各部分展示了对数据进行分类的过程。决策树通过对每个节点进行属性值的比较而得到分支，并在各叶节点得出分类结果。从决策树的根到叶的每一条路径就是对应的条分类规则，因此从决策树非常容易转换成分类规则。

决策树法在股票基本分析和技术分析中的模型思路大致是相同的，但是两者所需选取的变量是不同的。

1. 基本面分析的变量选取

采用个股财务报表中的流动比率、速动比率、资产负债率、销售毛利率、销售成本率、销售期间费用率、资产净利率、净资产收益率摊薄、主营业务利润率、营业利润率、股东权益率、净资产增长率、净利润增长率、主营业务利润增长率、主营业务收入增长率、营业利润增长、每股营业利润、每股主营业务利润、每股主营业务收入、每股资本公积金等自变量。另外，可以设定一个新变量“个股赢率”作为二元目标变量。若个股年累积收益大于流通市值加权市场年累计收益，则个股盈率为 1，相反则为 0。

值得注意的是，为了保证模型实证的有效性，基本面分析所选取的指标需要进行一定的筛选，筛选的条件如下。

(1) 研究期间所选取样本具备完整的财务报表数据。

(2) 研究期间所选取样本的停牌时间不得超过半年。

(3) 所选取样本近期未发生资产重组等影响模型有效性的重大事件。

2. 技术分析的变量选取

采用 20 日移动平均日收益方差、20 日移动平均日收益标准差、20 日移动平均流通市值加权日市场收益方差、20 日移动平均流通市值加权日市场收益标准差、20 日移动平均总市值加权日市场收益方差、60 日移动平均总市值加权日市场收益标准差、60 日移动平均日收益方差、60 日移动平均日收益标准差、60 日移动平均流通市值加权日市场收益方差、60 日移动平均流通市值加权日市场收益标准差、60 日移动平均总市值加权日市场收益方差、60 日移动平均总市值加权日市场收益标准差、换手率、日振幅、市盈率、波动率、日简单平均移动波动率、20 日简单平均移动波动率、60 指数加权移动平均波动率、成交量等自变量，并另外定义一新变量“下一日涨跌”为二元目标变量，若下一日涨则产生买点，设为 1，反之则产生卖点，设为 0。

对于决策树二进制的目标变量，主要有以下划分规则。

(1) 检验——Pearson 用于衡量目标变量，并依其建立分支节点。

(2) 熵值约简——通过对熵值大小的衡量反映节点不纯性，也称为熵不纯性。

(3) 基尼系数约简——通过对基尼系数大小的衡量反映结点不纯性，也称为 Gini 不纯性。

模型中叶节点的设定主要有节点最小观测数、叶子的最小观测数、节点最大分支数、决策树最大层数，如图 4.1 所示。

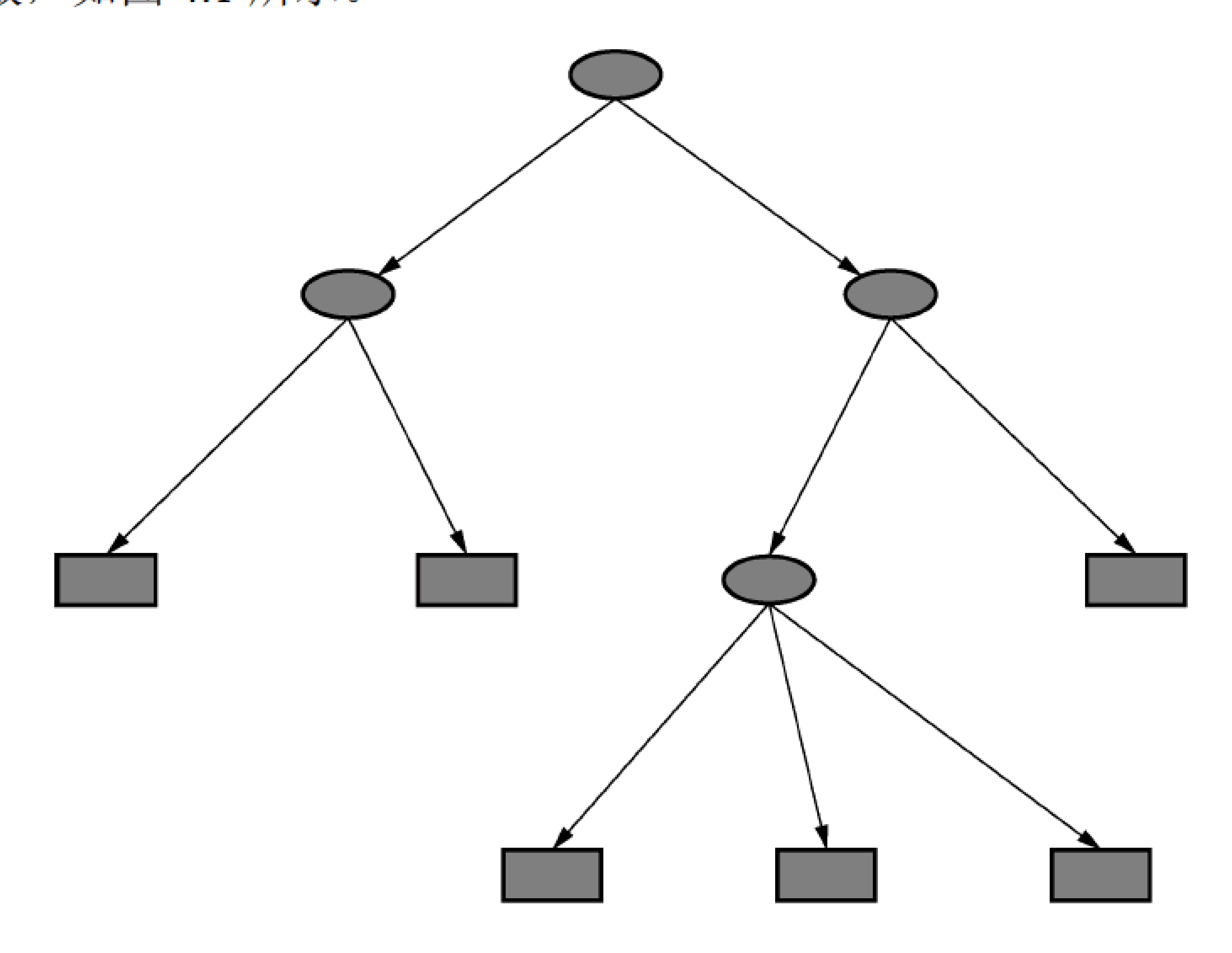


图 4.1　决策树模型结构

在成功建立模型后便可以得出决策树分类结果，得出训练数据与检验数据误差最小的叶子，根据分类规则可以判断每个筛选出的指标对因变量的显著性。并在此基础之上，便可以通过决策树模型对个股赢率/股票的买卖点进行预测。模型中对预测的结果也有正确率的估计。证券经营机构参考模型的预测结果和预测正确率，能够更好地确定投资策略、发布研究报告、提供投资顾问服务等。

4.1.4　聚类分析法的应用

聚类分析是对大量事先并未知晓其属性的数据集，按照数据的内在相似性将数据集划分为多个类别，使类别内的数据相似度较大而类别间的数据相似度较小。聚类分析能够作为一个独立的分析工具获得数据的分布情况，以便观察每一类数据的特征，集中对特定的数据群进行进一步分析。

应用到股票分析中，可通过提供一定的时间段，将该期间的股票数据进行分类，从中发现获利大而风险小的聚类，作为证券自营业务部门或者投资者的参考。

具体分析方法如下。

(1) 确定聚类分析的时间段、股票板块以及个股。

(2) 选取个股重要的财务指标进行分析，包括归属母公司净利润、每股收益摊薄、销售净利率、销售毛利率、营业成本率、期间费用率、销售费用率、管理费用率、财务费用率、营业利润率、成本费用利润率、应收账款周转率、总资产周转率、总资产收益率、净资产收益率摊薄、营业收入现金含量、销售现金比率、净利润现金含量等。

(3) 设计模型。

① 数据缺失值处理。

采用均值处理方式为 0 或为空的数值赋值。

② 数据标准化。

由于聚类分析对各变量间的数据规模差异十分敏感，因此将各变量转化为均值为 0、方差为 1 的新变量。

③ 模型参数设定。

将分类结果以 1，2…的形式标识为该聚类的 ID，以方便对结果的分析。同时设置最小平方数为聚类标准。另外，根据数据的具体情况确定最小聚类数以及最大聚类数。

(4) 得出实证结果。

建立模型后，模型会把数据分为几个聚类，并得到所有变量标准均值以及每一聚类变量标准均值的分布情况。通过对每个聚类具体情况的分析，可以确定哪一类适合进行投资。

① 若股票的净利润、净资产收益率、每股收益、营业利润率等衡量上市公司获利能力的指标均明显高于整个板块，而它的应收账款周转率、期间费用率、管理费用率、销售费用率等衡量上市公司运作成本的指标均明显低于整个板块，这类股票适合进行长期投资。

② 若公司总体获利能力的指标明显高于整个块，并且高于上述(1)情形，但它的营业成本率、财务费用率、应收账款周转率均低于该板块整体水平，说明该类企业获得巨额收益的同时也付出了大量的成本或者说所获收益短期内难以回笼。这类股票适合进行短期投资。

③ 若公司的总体获利能力以及上市公司的运作成本均低于整个板块，那么这类股票不适合进行投资。

4.1.5 人工神经网络算法的应用

人工神经网络是对人脑或自然神经网络的基本特性的抽象和模拟，它是通过模拟大脑的一些机理和机制，实现某种功能。具体来说，它是一组连接输入、输出单元，其中每个连接都和一个权重相关，通过调整这些权重，能够预测输入数据的正确类标号。根据网络的层次数，人工神经网络可以分为两层神经网络、三层神经网络和多层神经网络。其中最常用的就是三层神经网络。

这里的人工神经网络算法主要指的是误差反向传播算法(BP 算法)，它是一种监督式学习的人工神经网络，能将错误的讯号反馈回来，以便及时修正权重。BP 算法网络分为三层，分别为输入层、隐藏层与输出层，并通过转换函数的进行人工神经网络的网络训练，如图 4.2 所示。

(1) 输入层。接收外部环境输入信息，其处理单元即为变数个数。

(2) 隐藏层。为人工神经网络中最重要的部分，通常为一到两层，也可以没有隐藏层。

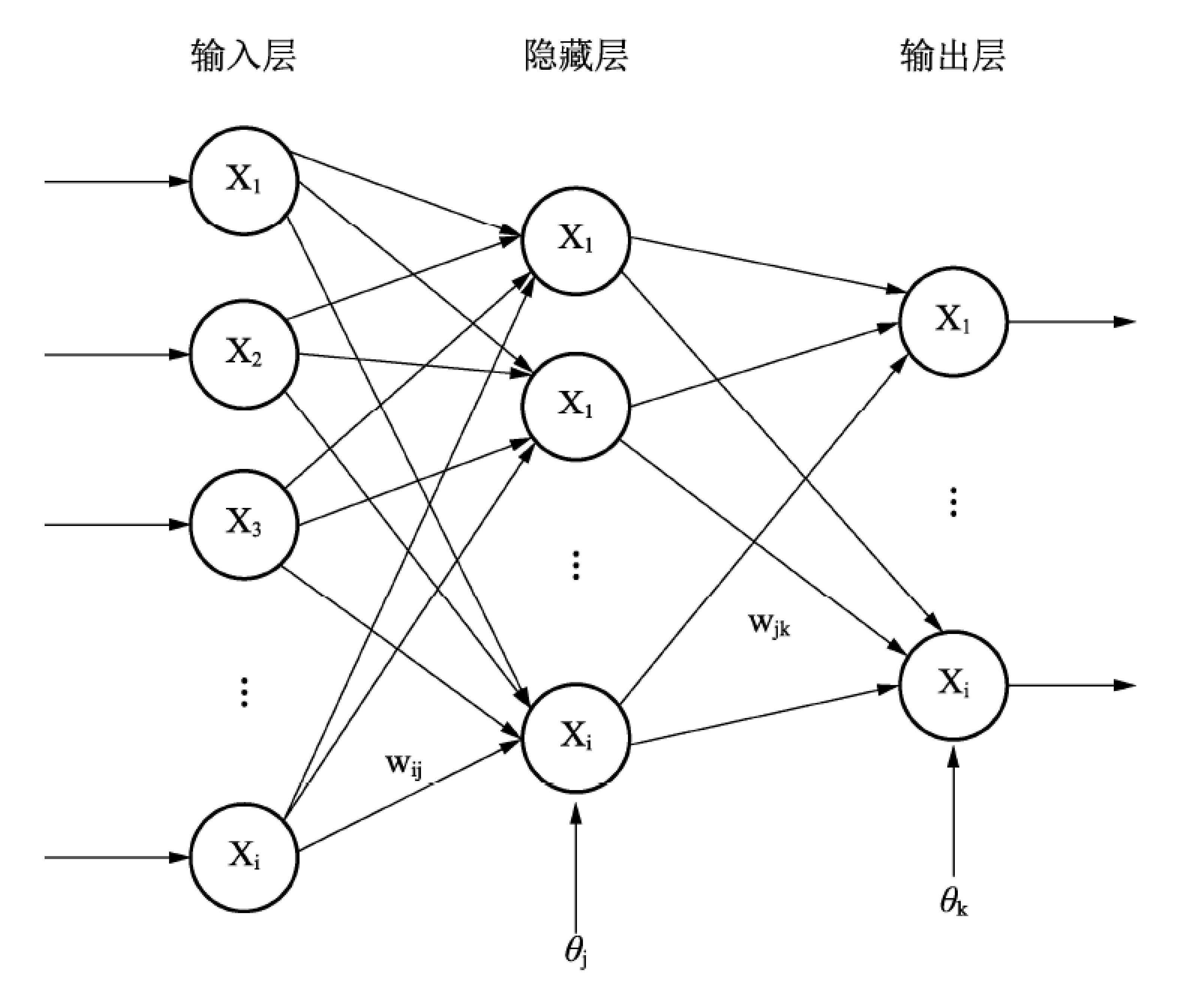


图 4.2 BP 算法流程

(3) 输出层。输出网络的处理结果，输出层的处理单元数为研究问题所要得到的结果个数。

1. 基于人工神经网络算法的股票基本面分析

1) 样本选择

选择上市股票一段时间的一定数量股票的发展能力数据和季度报酬率数据。

2) 变量选取

将上市公司发展能力的指标作为输入变量，包括主营业务收入增长率、营业利润增长率、利润总额增长率、净利润增长率、净资产增长率、流动资产增长率、固定资产增长率、总资产增长率、摊薄每股收益增长率、每股净资产增长率等 10 个指标。另外，将个股赢率作为输出变量，并把它分为 3 类。第一类(个股赢率>0.052)代表个股当季表现优于大盘；第二类(−0.092≤个股赢率≤0.052)代表个股当季表现与大盘相当；第三类(个股赢率<−0.092)，表示个股当季明显劣于大盘。

3) 模型建立

(1) 将输入变量做标准化处理，其公式为

$$Z = \frac{X - U_x}{\sigma_x} \tag{4.1}$$

(2) 隐藏层。

隐藏层中相关参数包括采用神经元个数、层数、激活函数、联合函数共 4 项，说明如下。

① 层数。该参数值的决定方法各不相同，但大部分学者均认为该值为 4 以下并不会影响神经网络的训练结果。因此，隐藏层根据具体情况可以设置为 1、2 或者 3 层。

② 神经元个数。一般认为，神经元个数应大于等于输入变量个数和输出变量个数的最小值，小于等于输出变量和输出变量个数之和，在建模过程中根据上述规则设置合理的参数范围，并采用试凑法进行逐一尝试。

③ 激活函数。常用的激活函数有 Logistics 函数、双曲正切函数、反正切函数、高斯函数，Logistics 函数适应性更强，故一般将其作为激活函数。

④ 联合函数。联合函数分为曲线连接和线性连接，一般以线性连接为主。

(3) 输出层。

在输出层中主要有输出变量、误差和误差函数 3 项参数。

① 输出变量：由于前面将个股赢率分为 3 个类别，因此根据这 3 个类别所对应的分布设计为 3 个输出变量，取值在 0 到 1 之间，用其作为评判股票季度表现的指标，当预测值落在该类别的概率大于 1/3 时，说明个股属于该类别的可能性较高，概率越大可能性越高。

② 误差：误差计算方式为输出处理单元与目标值之间差异平方和的一半，其公式为

$$E(w)=\frac{1}{2}\cdot\sum(d_l-y_l)^2 \tag{4.2}$$

其中，w 代表网络中所有权重之和；d_l 代表第一个输出神经元的实际值，即个股的真实赢率；y_l 代表第一个输出神经元的预测值，即个股的预测赢率。

③ 误差函数。

误差函数主要有伯努利函数、多重伯努利函数、Logistics 函数、柯西函数等。在股票基本面分析中，多重伯努利函数更适用。

4) 建模过程

(1) 将网络结构设置为 Multilayer Perception(MLP)结构，模型标准设置为 Profit/Loss，即设置了损失矩阵的模型。

(2) 将收集的数据的 70%作为训练数据，得出历史个股赢率的统计。

(3) 用试凑法逐一测试选出最小误差，得出最优节点。

(4) 选取一定的股票最近期间的发展能力指标为输入变量，对其下一季度的个股赢率进行预测，得出预测结果。

2. 基于人工神经网络算法的股票技术分析

人工神经网络算法应用于技术分析，是将股票的技术指标作为分析变量，通过分析一段历史时期内的股票技术指标的变动来预测股票未来走势。其选用的模型和模型设计过程与上述基本分析相类似，主要不同在于变量的选取以及预测方面。

在变量选取方面，一般选取个股研究期间内的短期技术指标作为变量，主要是指：每日 9:30 至 15:00 期间每五分钟涨幅、每日涨幅、每日振幅、每日成交量涨幅、每日上证综指涨幅、隔日开盘涨幅等 6 个指标。

此外，将下一日的开盘价涨幅作为目标变量。下一日开盘价涨幅=(前一日收盘价−下一日收盘价)÷前一日收盘价。

预测方面，与基本面分析不同，技术分析主要分为 3 个部分，具体说明如下。

(1) 输入预测数据集。将预测样本输入数据集，并设置隔日开盘涨幅为目标变量。

(2) 打分。将 3 个层的训练过程进行打分，并将得分代码应用到预测数据集。

(3) 预测。对预测时间段的隔日开盘涨幅进行预测值与实际值的比较。

最终通过建立拟合方程，得出预测值与真实值之间的关系，从而为投资决策提供参考。

4.2 客户关系管理

客户关系管理(Customer Relationship Management，CRM)是一个获取、保持和增加可获利客户的方法和过程，通过提高客户的忠诚度而最终提高企业利润率。

证券公司通过实施客户关系管理，提供快速、周到的优质服务，可吸引和保持更多客户，从而提高核心竞争力。要做好客户关系管理，证券公司应当利用大数据技术对客户的信息做深入的分析，做好客户细分，为不同的客户提供个性化服务。同时也要对流失客户进行科学的分析和预测，使证券公司能够尽早提出相应措施，避免客户流失或者使客户流失最小化。

4.2.1 客户细分

国内证券公司拥有大量的客户群，客户多种多样，对于不同的客户，他们的需求也有所不同。证券公司受自身条件的限制，不能同时满足所有客户的需求，因此采取客户细分策略对于证券公司优化资源配置、证券公司内部管理、实现券商价值最大化都起到了至关重要的作用。

1. 证券客户细分的作用

第一，对客户进行细分，设置相对应的客户级别。筛选出其中最有价值的客户，并且针对这些客户采取个性化服务，有助于提高客户的忠诚度与满意度。

第二，有助于证券公司探索到新的市场机会。

第三，有助于证券公司研发新的金融产品，以满足客户的需求。

第四，有助于证券公司挖掘高净值客户，加强对高净值客户的抢夺力度，提高公司竞争力。

2. 客户细分模型

1) 客户等级体系模型

为了在客户细分的研究中建立完整的功能结构模型，在此基础上建立一套标准的客户细分模型——客户细分的 DFM 模型。该模型包括数据(Data)、功能(Function)及方法(Method) 3 个部分，因此将此模型命名为客户细分的 DFM 模型，如图 4.3 所示。

目前，简单的传统的客户信息模型体系已经不能适应证券公司开发和营销集合理财产品和服务的需求。因此，应当将数据挖掘技术应用于证券客户资料中，发现隐藏其中的规律，建立符合实际情况的客户等级体系模型。客户等级体系模型建立的过程如图 4.4 所示。

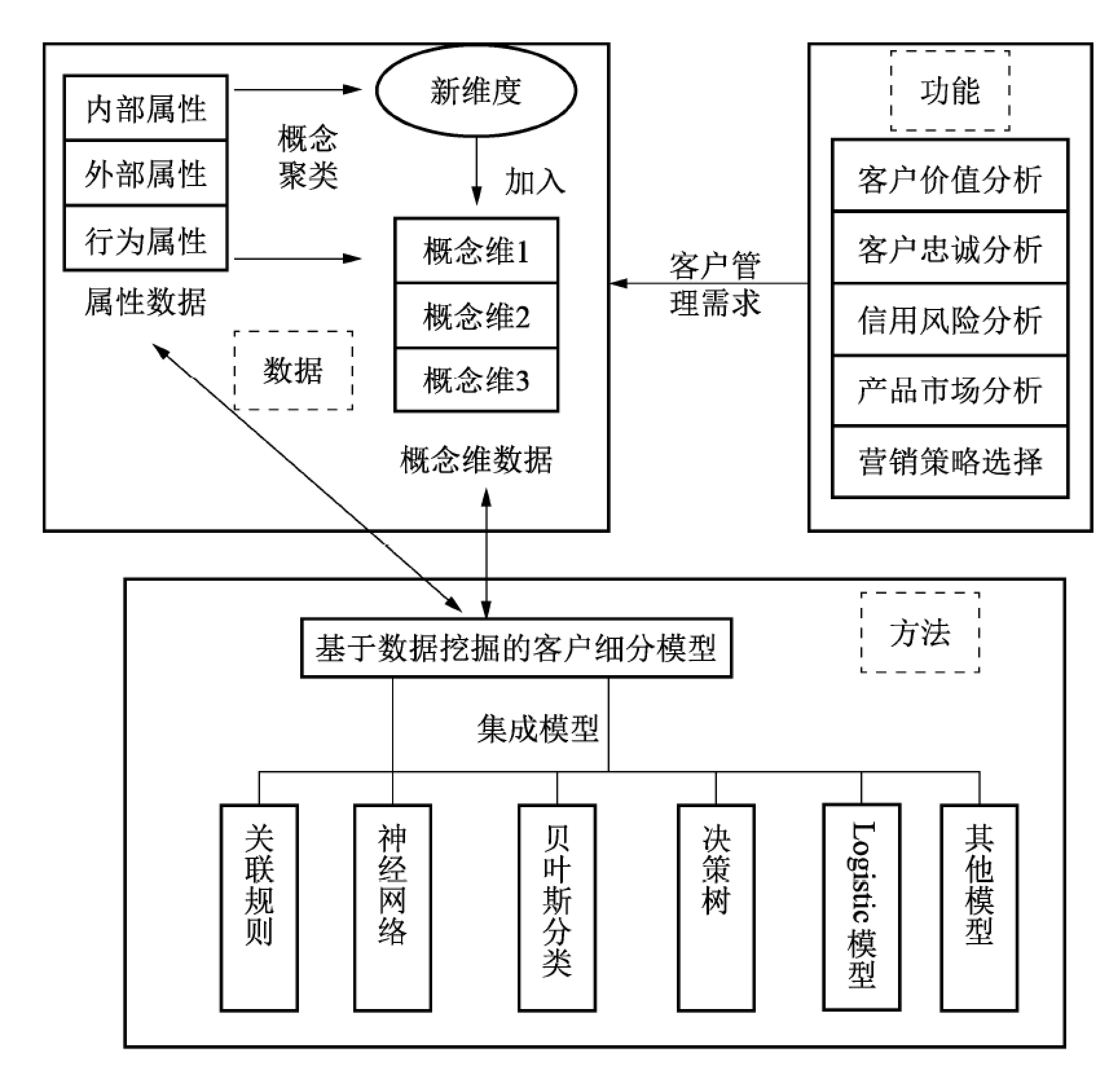


图 4.3　客户细分的 DFM 模型

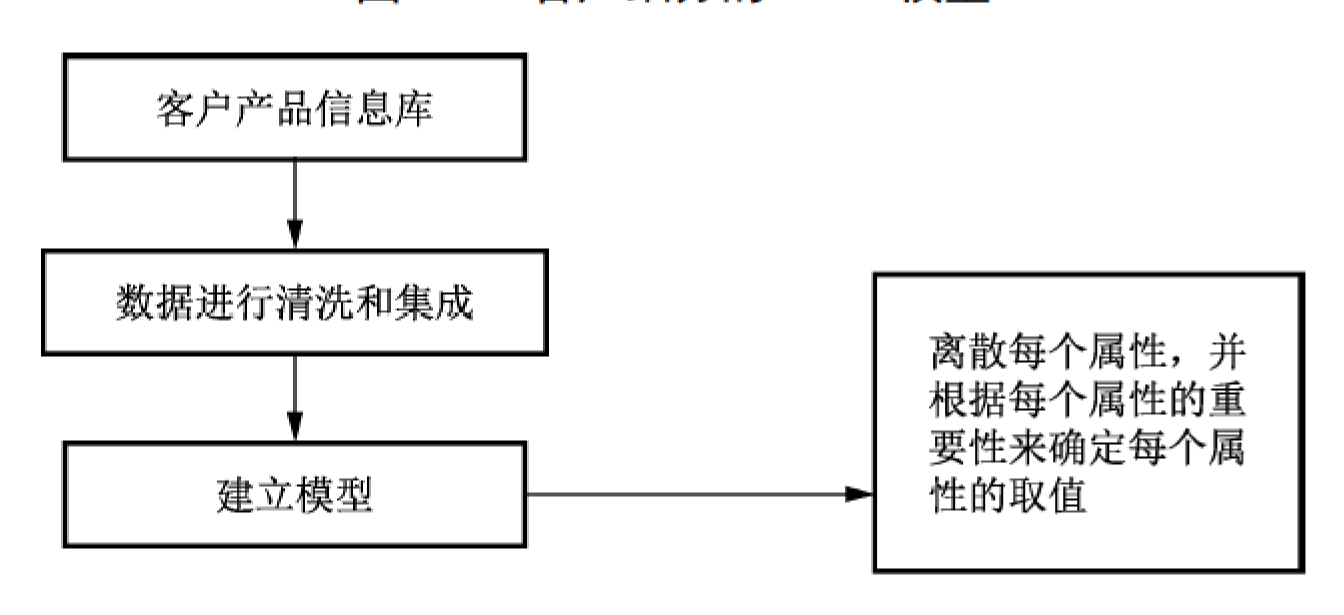


图 4.4　客户等级体系模型建立过程

评定客户的等级，必须先获得客户的基本信息情况，如性别、年龄、收入状况、信用状况、投资偏好、风险偏好等多个项目，其后经过数据清洗和集成，经过数据转换，最终得到基本的客户信息数据表(见表 4.3)。

表 4.3　客户基本信息数据表

性别	投资风格
年龄	风险偏好
婚姻状况	投资广度
最高学历	证券账户资产

续表

职业	资产周转率
行业	仓位
财产收入	盈利率
总资产(区间)	交易活跃度

一般情况下，证券公司会以客户证券账户资产以及交易活跃度作为评定客户等级的主要标准，对不同等级的客户其服务策略也不同(见表 4.4)。

表 4.4　客户细分及其服务策略

客户细分类型	交易行为特征	服务策略
类型一：主要客户	资产总量不大，有一定量的买卖交易和资金存取；无专业知识，投机性较强；数量较多，大多数为中小散户，佣金贡献成交量大	提供大众性咨询服务(如开办讲座等)，提高其投资能力
类型二：睡眠客户	资产总量很少，基本不进行股票交易，也不存取资金；有少量的收益；数量不多	不必对该类客户进行关注，尽可能地减少该类客户数量
类型三：优质客户	资产总量大，交易操作频繁；有专业的投资知识，对市场非常敏感，有较高的盈利能力；是公司利润的主要贡献者	应经常与客户沟通，及时发现客户的真正需求，保持客户的满意度与忠诚度
类型四：潜力客户	有一定的资产量，交易操作次数少，现金存取频率低；收入稳定，投资渠道少；对市场不敏感，盈亏不大；对公司的佣金贡献量不大	可以推荐一些信托产品或者代客理财，或者提供投资咨询服务，改变其投资观念，将其发展成为优质客户

2)　证券客户价值分类模型

国内证券公司在客户分类方面方法较为简单，并不能很好根据客户的需求和特点划分客户群。国外的证券公司采用的是 SOM(Self Organizing Map)聚类分析方法对客户价值进行细分。这种方法值得借鉴。

客户价值是指企业在与客户的交往过程中，从客户那里获得的客户总价值与企业支付的总成本的差额。国内证券公司目前主要应用客户的资产、交易量、贡献度等统计信息进行客户价值细分。从客户价值上把客户细分为：高利润客户、次级利润客户、低或无利润客户。

SOM 神经网络是较为广泛应用于聚类的神经网络，它是一种无监督学习的神经元网络模型。SOM 网络可以采用各神经元之间的自动组织去寻找各类型间固有的内在的特征，从而进行映射分布和类别划分。所以神经网络对于解决各类别特征不明显、特征参数相互交错混杂的、非线性分布的类型识别问题是非常有效的。

在 CRM 系统中应用客户价值分类模型主要分为以下几个阶段：收集证券客户数据、数据预处理、客户聚类、模型评估、证券客户基于客户价值的分类，如图 4.5 所示。

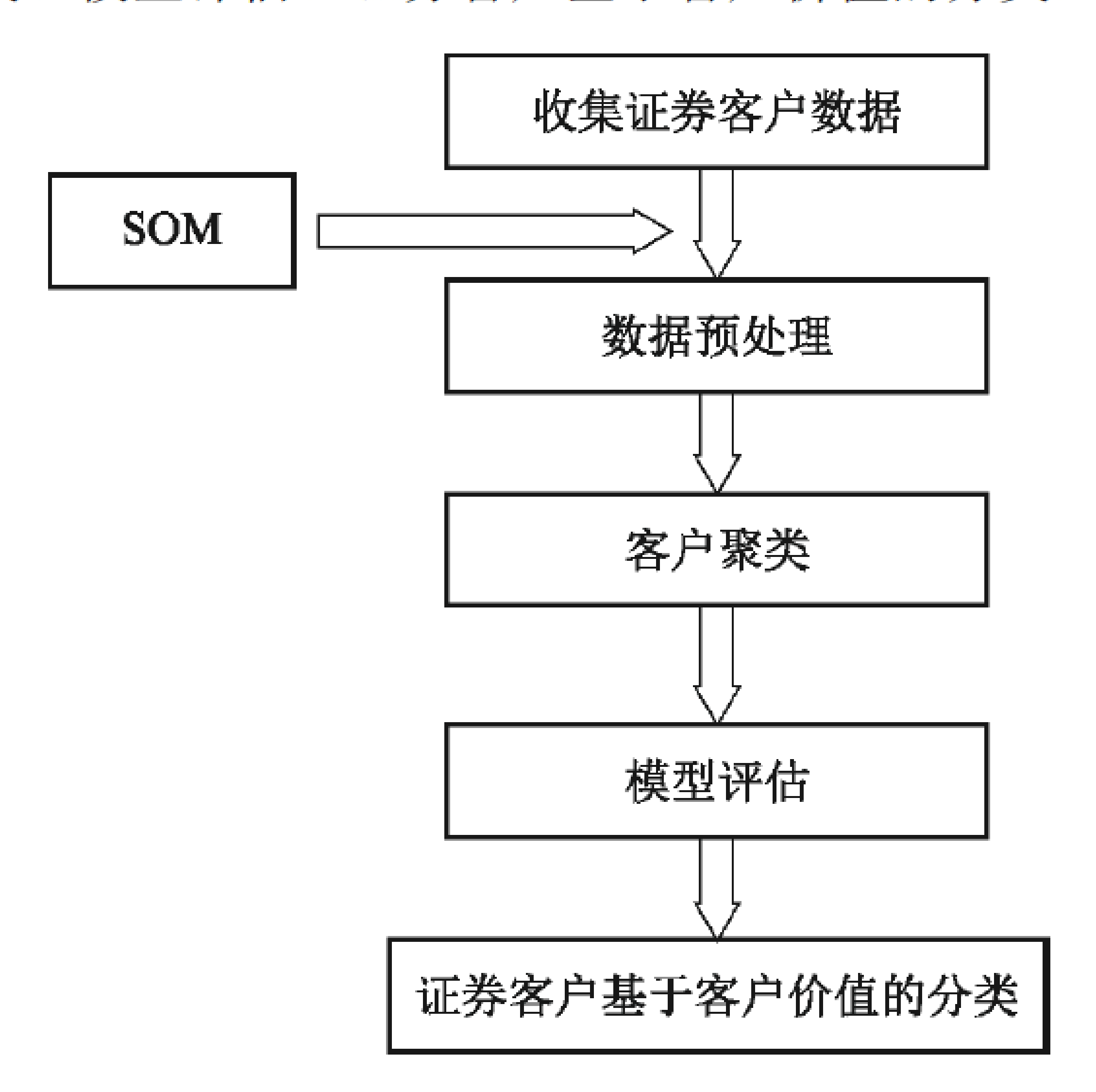


图 4.5 基于 SOM 聚类分析的证券客户分类流程

4.2.2 客户满意度

客户满意度(Customer Satisfaction Research，CSR)，是指客户的期望值与客户实际体验的匹配程度，换句话说，就是客户通过对一种产品或服务可感知的效果与其期望值相比较后得出的指数。

客户是企业的核心资源，如何让客户满意证券公司提供的服务或产品成为证券客户关系管理的一个十分重要的分析方面。证券公司有必要设计客户满意度评价指标体系从而对客户满意度进行研究，同时为挖掘潜在客户、提高客户价值、提高客户的满意度提供技术上的支持。因此，对证券公司来说，构建客户满意度模型十分重要。

1. 客户满意度关系模型假设

第一，假设证券客户满意度是由客户对实际感知的服务和期望服务质量之间的差额决定的。

第二，证券客户的满意度是客户对服务价值的一种评估。

第三，证券客户总的满意度主要受客户对证券公司的服务或产品的评价影响。

第四，客户对各个服务或产品的评价之间是相互独立的。

2. 证券客户满意度关系模型的建立

根据以上 4 个假设，再加上给定的客户满意度评价指标，并结合证券公司管理的特点，分析影响证券客户满意度的因素，构建证券客户满意度关系模型(见图 4.6)。

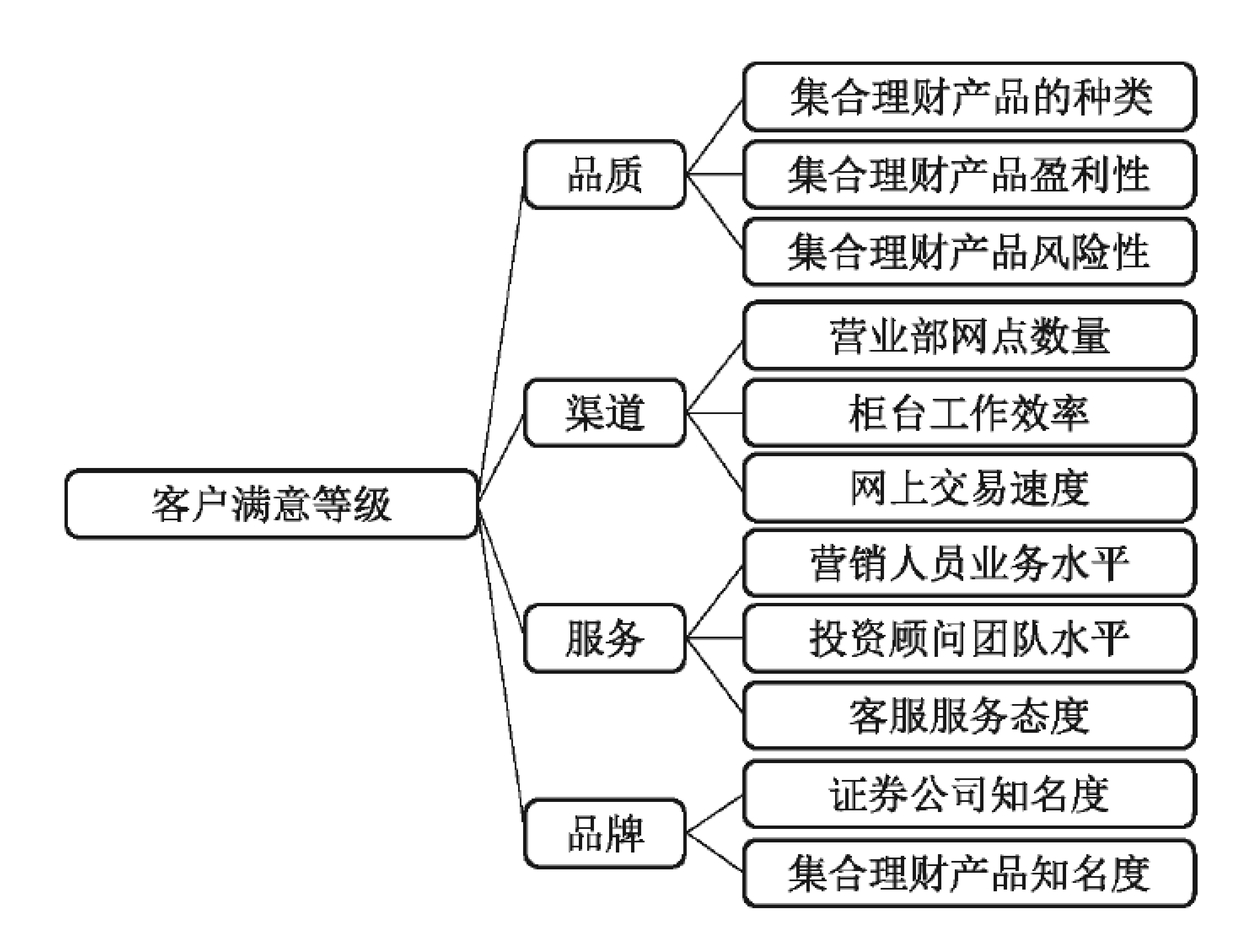


图 4.6　客户满意度评价体系

3. 客户满意度衡量与分析

客户的满意度的提升或下降是由客户实际感受到的服务质量与客户预期的服务质量之间的比较决定的。

用函数关系来表示：客户满意度=f(客户期望值 E，实际获取值 A)

即：客户满意度=客户实际获取服务质量数值/客户期望服务质量数值= A/E

客户满意度由此也可以划分为 4 种情况：客户满意度>1、客户满意度=1、客户满意度<1、客户满意度<0(见表 4.5)。如图 4.7 所示为证券客户满意关系模型图。

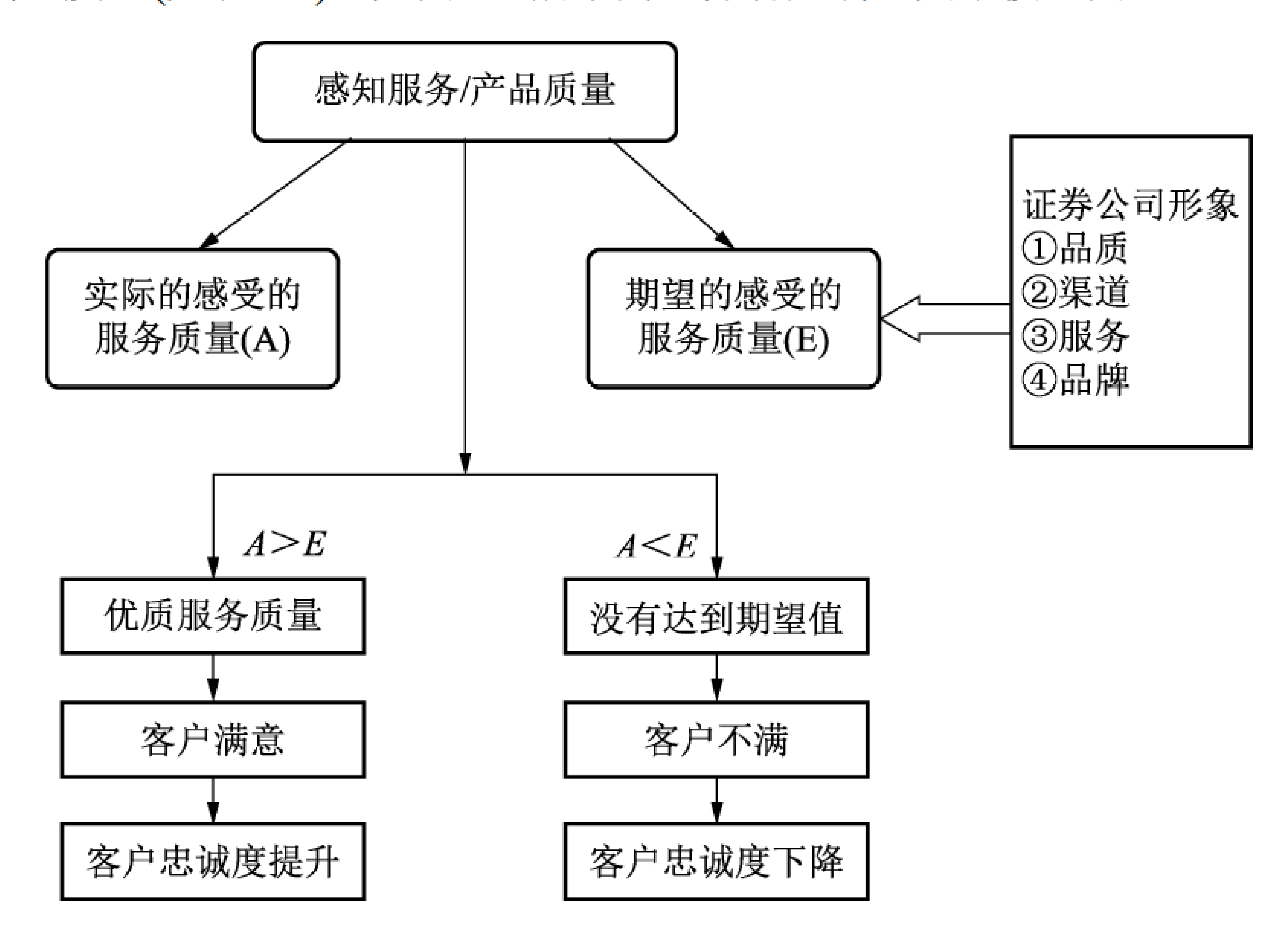


图 4.7　证券客户满意关系模型

表 4.5 证券客户满意度测评表以及对应策略

评分	客户满意程度	客户忠诚度	采取对应策略
客户满意度>1	愉悦	非常忠诚	继续挖掘客户期望，提供优质的个性化服务或产品
客户满意度=1	满意	比较忠诚	加深客户服务或产品，防止客户流失
客户满意度<1	不满意	客户容易转户	提高服务和产品基本成分的可靠性，尽可能地减少客户流失
客户满意度<0	强烈不满	客户转户销户	—

从证券公司角度看，要增加客户的满意度，有以下几个途径。

(1) 降低客户的期望值。

(2) 提升客户的实际获取值。

对此，证券公司应该采取以下措施。

1) 降低客户期望值

降低客户期望值是比较困难的，尤其在证券业中，客户的需求是多样的，因此，导致客户的期望也是多种多样的。那么，证券公司应该充分利用客户期望值的脆弱性来降低其期望值。比如证券客户对于以往的投资经验或习惯刚建立起来的期望值通常是非常不稳定的，只要稍微受到市场波动就会产生动摇，这时候就应该抓住机会，通过公司专业的投资顾问团队提供给客户其他的期望值，送样就可以降低客户的期望值。

2) 提升公司的服务水平和产品质量

提高客户实际获取值是企业通常的做法，证券公司也不例外。通过证券公司分类模型把客户进行细分，分成高价值客户、中价值客户、低价值客户，随后根据不同的客户提供不同的服务或产品，最终客户产生满意的感觉。

4.2.3 流失客户预测

当下证券行业竞争十分激烈，券商之间存在严重相互抢夺客户、客户流失的现象。客户是企业的核心资源，衡量一个证券公司的成功关键在于客户，客户的投资收益、客户份额等都与证券公司利润密切相关。据推算，挖掘一个潜在客户并最终使他成为正式客户是留住老客户成本的 6～7 倍。通过创建客户流失模型进行预测，可以使证券公司做出相应的预防措施，从而避免客户流失抑或使客户流失最小化。

1. 证券公司客户流失的原因

证券公司客户流失的原因是多方面的，可以分为以下几类。

1) 自然流失

客户自然流失是因为营业部的搬迁、撤销等。这种情况是不可避免的，这种客户流失是在证券公司可承受范围之内的，不具有持续性。

2) 竞争流失

竞争流失就是因为各个证券公司之间的竞争导致的客户的流失。

例如，其他券商在佣金上具有优势，投资顾问团体投资咨询水平更高等。这些因素都可能导致客户流失。

3) 过失流失

过失流失是由于客户对证券公司的服务质量产生不满意而造成的。例如，客户经理服务态度不好，不能满足客户正当需求等。

2. 客户流失建模的原则

客户选择哪家证券公司，选择购买哪种金融产品和服务，客户的诸如此类的选择会受到各方面因素的影响。从微观角度来说，与人情关系、价格、服务质量、竞争对手的策略有关；从宏观角度来说，国家政策、国际形势的变化都会对客户流失与否造成一定的影响。在这种情况下，要非常精确地预测某个客户的流失是无法做到的。但是，大多数情况下客户的行为是理性的，他们不会随意离开目前的证券公司，并且客户流失之前都会有一些相似的行为特征，这就使预测客户流失的做法成为可能。

3. 客户流失预测模型的建立

客户流失对于证券公司是一个非常严峻的问题。证券客户流失不仅仅是指客户销户情况，也是指客户把大量资金转出证券公司购买其他金融机构的理财产品或信贷产品等。因为面对投资趋于多样化，证券市场持续低迷，导致客户投资股票的意愿大幅降低。然而，证券公司应用数据挖掘技术进行证券客户流失分析具有十分重要的意义。建立证券公司客户流失预测模型，可以了解到哪些客户会流失，客户最近有哪些异常行为，还可以分析客户流失的原因。通过这些现象的分析，证券公司就能在客户流失前采取相应措施。因此，创建客户流失预测模型具有重要的现实意义。

建立客户流失模型预测的具体流程如下。

1) 确定业务问题与环境评估

将客户流失分为自然流失、竞争流失、过失流失 3 类。

2) 数据收集与处理

为了建立损失客户预测模型，必须寻找大量的原始数据，然后对数据进行简单的处理，随后再将数据转换成模型。

在建模时，要根据数据的特征并对数据进行分析以寻找出不同数据之间的关联度，寻找出哪些变量与客户流失有关，哪些与客户流失无关。从而排除无用数据，降低模型的复杂性，使模型预测更加精确。

3) 数据应用和评估

在损失客户预测模型建立后，需要大量数据进行反复检验。如果数据检验与预估数值相同就可以立即运用到当前业务中。通过模型预测客户流失的趋势采取相应措施；反之，如果预估数值存在很大偏差就构建新的模型，如图 4.8 所示。

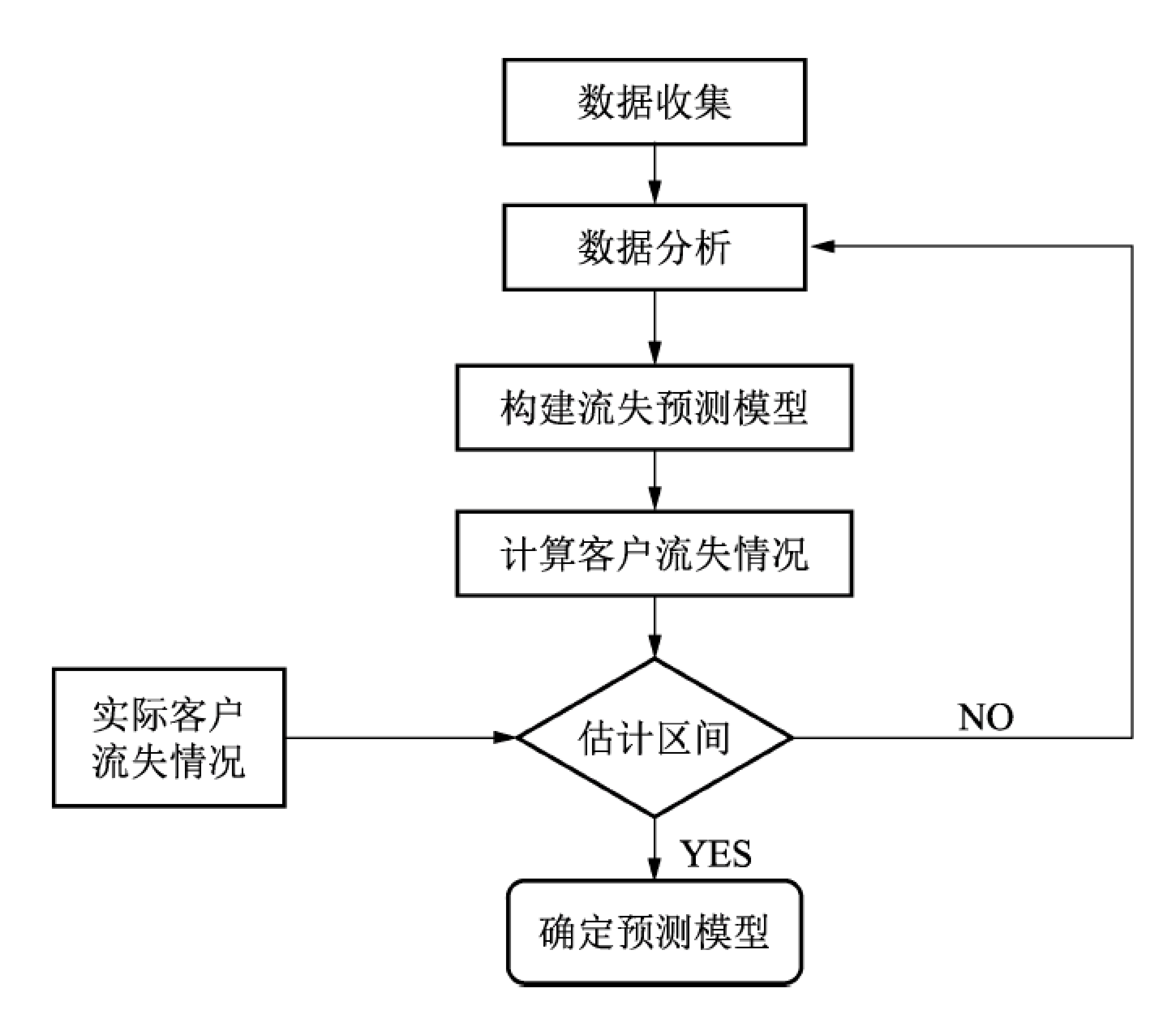


图 4.8　损失客户预测流程

4. 客户流失的 Logit 模型

一般地，损失客户预测模型是通过客户数据资料的分析和研究建立 Logit 回归模型。利用此模型发现客户的异常行为，提前做出相应措施，防止客户流失。

1)　变量选择

建立具体的客户流失预测模型的关键是恰当地选择影响客户流失的变量，即建模变量。

影响证券公司客户流失的主要变量有：证券客户资产、集合理财产品、异常大额交易、异常银证转账、客户投诉情况等。

2)　通过交叉表来判别显著影响变量

通过运用 SPSS 等工具对上述变量进行研究分析。交叉表分析结构主要有 3 个指标，“流失百分比”“全部百分比”“比率值”(见表 4.6)

表 4.6　各属性识别流失客户的情况

属性	流失百分比	全部百分比(%)	比率值
			/
			/
……	……	……	……
			/

3)　建立 Logit 模型

假设个体选择方案 i=1(客户不流失)的概率为 P，则 i=2(客户流失)的概率为 $1-P$，记为 Q。那么 P、Q 与影响因素之间的关系用以下模型表示：

$$\begin{cases} P = \mathrm{e}^z / 1 + \mathrm{e}^z \\ Q = 1/1 + \mathrm{e}^z \\ z = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \\ \mathrm{Logit}(P) = \ln\left(\dfrac{P}{Q}\right) = Z = \beta_0 + \beta_1 X_1 + \cdots + \beta_n \end{cases}$$

其中，β_0 是与各因素无关的常数项，β_1、$\beta_2 \cdots \beta_n$ 是回归系数。Logit 分析法可以考察多个变量对证券客户流失的影响，能够通过对每一个属性变量进行分析研究以考察它们的贡献程度，然后淘汰一些不重要的变量，最终选择贡献度最大的属性变量进行 Logit 回归分析。

证券公司能够识别将要流失的客户意味着能够减少维护客户的成本，这意味着增强客户与公司之间的关系。此外，证券公司还应该注意到随着证券新业务的发展，在根据原有的数据库信息构建和推导的客户流失预测模型的效果可能随着时间的变化而逐渐衰退，模型需要不断地更新和改进。从客户管理角度来说，维护和正确使用数据库是十分重要的。

4.3 投资情绪分析

在实践中，投资者的非理性行为在一定程度上会影响金融市场(例如 2015 年中国中车的大起大落)。在理论上投资者是否理性是传统金融学和行为金融学的分水岭。传统金融理论认为投资者是理性的，并没有考虑到投资者的情绪因素，而行为金融理论认为投资者易受到情绪、情感等因素的影响，并将投资者情绪作为其两大基本假设之一。

投资者情绪是一个模糊和非数量化的概念。从广义上看，投资者情绪包含诸多能够影响投资者的证券估值和市场预期的因素；从狭义上看，投资者情绪仅研究对投资者的证券估值和市场预期能够产生影响的经济变量和其他因素。对于证券经营机构与相关研究机构来说，投资者情绪的测量是一个难题。如何对投资者情绪进行量化分析，这对股票市场研究来说至关重要。

4.3.1 投资者情绪的测量

关于投资者情绪指标，根据指标数据的主客观性和数据来源可以分为两类：主观测量指标和客观测量指标。近年来，对于投资者情绪的测量出现了新的趋势，学者们针对主观测量指标和客观测量指标的不足，在其基础上加以改良，构造了复合投资者情绪指标作为投资者情绪的代理变量。下面分别加以介绍。

1. 主观投资者情绪指标

主观投资者情绪指标也称为直接指标，是指经过调查得到的直接反映投资者对市场行情的看法和判断，一般以投资者看涨、看跌及看平的比率数据来表示出来，或是用经济信心指数进行替代，直观地表现出投资者对未来市场的悲观或乐观情绪。例如：美国证券市场的友好指数、个人投资者协会指数(AAII 指数)、投资者智慧指数(II 指数)等。由于国内

尚无与投资者情绪调查有关的标准化组织，国内一些机构编制出的投资者情绪调查指标尚缺乏权威性。在研究中少有学者运用主观指标来度量中国投资者情绪。就目前而言，具有一定影响力和认同度的主观指标主要有央视看盘指数、消费者信心指数、巨潮投资者信心指数以及耶鲁-CCER 投资者信心指数

1) 国外常见的投资者情绪指标

(1) AAII 指数。

AAII 指数，即美国个人投资者协会指数，它是由美国个人投资者协会自 1987 年调查发布的指数。AAII 指数每周通过随机抽样向其会员发出调查问卷，并于周四记录当周收回的问卷。调查的内容是要求参与者对未来 6 个月的股市进行预测：看涨、看跌或者看平。由于调查主要针对个人，所以该指标一般用以衡量个人投资者情绪，也是国外学术界常用的投资者情绪指标之一。

(2) II 指数。

II 指数，即投资者智慧指数，它是由 Chartcraft 公司编制的一个对超过 150 家报纸股评人士情绪的调查数据。它的具体公式为：看多比例与看空比例之差。由于股评的作者大多都是现任的或者是已经退休的金融专业人士，他们具有一定的专业性，因此 II 指数被视作中型投资者的情绪的代表。

(3) 友好指数。

友好指数是美国哈达迪(HADADY)公司的产品，于每周一在美国证券交易所闭市后公布。该公司统计全国主要报刊、基金公司、投资机构等每周的买进卖出建议，然后通过打分评估它们的乐观程度。

2) 国内常见的投资者情绪指标

(1) 央视看盘指数。

央视看盘指数由中央电视财经频道编制，通过向知名的机构投资者和普通个人投资者发放调查问卷，收集其对后市的看法编制而成。问卷中将投资者对市场的预测分成看涨、看平和看跌 3 类，调查分为日调查和周调查两种。

(2) 消费者信心指数。

主观指标中有一类是使用其他经济信心指数来替代，我国的消费者信心指数也常用于作为投资者情绪的代理变量。消费者信心指数由国家统计局编制，用以衡量社会公众对目前及未来经济的信心程度，在一定程度上能反映投资者情绪。

(3) 巨潮投资者信心指数。

2003 年，深圳证券信息公司借鉴国外已有的投资者信心指数及国家统计局的消费者信心指数的编制方法，推出了巨潮投资者信心指数，它由一组动态的量化指标构成，刻画了投资者对目前及未来市场的信心状态，各指标数据均由每周一次的问卷调查获得。

2. 客观情绪测量指标

客观指标也称为间接指标，这类指标主要是采集金融市场上与投资者情绪相关的公开交易数据或通过相关的统计方法来构造相应的情绪指标来衡量投资者情绪的变化。相对于主观指标而言，客观指标在学术研究中应用更为广泛。早在 20 世纪 80 年代，西方学者已

开始收集证券市场上与投资者情绪有关的数据，并对这些数据进行处理并构造相应的情绪指标，作为投资者情绪的度量。这些指标根据其来源与性质的不同，可以大致分为以下四大类。

1)　市场表现类

市场表现类包括：腾落指数、新高新低指标、首日 IPO 表现(包括 IPO 发行数量和 IPO 首日收益)、市场换手率、市场交易量、市场流动性水平。腾落指数是以股票每天上涨或下跌的家数作为观察与计算的对象，以了解股市人气的盛衰，研判大盘的走势。新高新低指标是计算市场上的股票创一年来新高或新低的家数，以此反映市场的强弱程度。IPO 市场相关指标包括了 IPO 发行数量和 IPO 首日溢价。一般认为一段时间内 IPO 发行数量越大，首日 IPO 溢价越高，投资者情绪越乐观。市场交易量和市场换手率指标也都是常见的客观投资者情绪测量指标，市场交易量越大，换手率越高，市场中的投资者情绪越乐观。流动性水平也常出现在国外关于投资者情绪的研究中，作为投资者情绪的代理变量。

2)　交易行为类

交易行为类包括：保证金借款比例、短期利率变化比例、卖空比例、零股卖空比例。在交易行为类指标中，美联储每月发布的保证金借款比例常被认为是牛市指示器，保证金借款比例越高，市场中投资者情绪越乐观。短期利率的变化常被看作是熊市指示器。卖空比例是卖空交易额占总的卖出交易额的比重，卖空比例越高，投资者情绪越悲观。零股卖空比例是代表着不足 100 股的买卖交易占总交易额的比例，零股卖空比例更多地反映的是个人投资者情绪，零股买卖比例越高，投资者情绪越悲观。

3)　衍生变量

衍生变量包括：认沽认购比、波动率指数 VIX。认沽认购比代表着卖出/买入期权的交易量之比，该比例越高，代表着投资者情绪越悲观，常作为熊市指示器。波动率指数 VIX 又称为“恐慌指数”，用以反映 S & P500 指数期货的波动程度。

4)　其他情绪代理

除以上 3 类指标外，还有一些能反映投资者情绪的客观指标，包括封闭基金折价率、共同基金净买入、红利溢价、新增投资者开户数、股票发行/债券发行比例、季节性情绪变化 SAD 等。

证券经营机构和研究机构可以根据上述投资者情绪指标，以 SPSS 统计软件为工具，建立一个关于投资者情绪指标与股票市场价格之间关系的模型，从而为投融资服务客户提供一定的参考。

4.3.2　基于网络舆情的投资者情绪分析

1. 网络舆情与投资者情绪

投资者情绪除了表现在上述已经被量化的指标上外，还会在网络舆情中体现。随着互联网的普及，以微博、论坛、博客等为代表的网络社交媒体广泛流行，网络舆情逐渐成为影响人们情绪、态度行为的重要因素。

网络舆情(Network Public Opinion)，是指在互联网上流行的对社会问题不同看法的网

络舆论，是社会舆论的一种表现形式，是通过互联网传播的公众对现实生活中某些热点、焦点问题所持的有较强影响力、倾向性的言论和观点。它具有以下几种特征。

1) 直接性

直接性是指网民可以通过微博、论坛和博客随时发表意见，民意表达十分畅通；网络舆论具有无限次即时快速传播的可能性，网民可以转发将信息重新传播，一个爆炸性的新闻信息能在很短的时间被大多数网民获取。

2) 虚拟性

互联网由于是一个虚拟的空间，发言者的身份是隐蔽的，再加上我国对网络舆情的管理和监督不够完善，因此网络舆情的真实性是值得推敲的。有的信息可能是网民片面、错误的认识，有的信息可能是网民宣泄情绪所捏造的，有的信息也可能是出于商业目的甚至是不法目的杜撰的。因此，网络舆情具有一定的虚拟性。

3) 突发性

网络舆情的形成往往非常迅速，一个新闻热点再加上一个情绪化的观点就可以掀起一大片舆论的波浪。

4) 随意性和多元性

网络舆情不同于传统媒体的一点是网络舆情对个人来说是没有门槛的，所有人都可以通过网络媒体发表意见和评论。网民在网上或隐匿身份、或现身说法，谈论国事、交流思想。网络为民众提供交流的空间，也为搜集真实的舆情提供了素材。

在金融领域，越来越多的投资者会在网络中表达自己的投资情绪，同时投资者的投资决策会受到网络舆情的影响。网络舆情中的投资者情绪对证券经营机构来说具有极高的研究价值。原因如下。

首先，投资者情绪会对股票价格产生系统性影响。当投资者情绪好的时候，投资者倾向于采用简单启发式来辅助决策，并在信息处理中较少采取批评的模式；而在情绪不好的时候，投资者更倾向于采用更加周密的分析活动，但是投资者通常会将自身的情趣归咎于错误的来源而产生错误的判断。投资者的个体情绪变化可以通过网络媒体在群体中蔓延传染，最终会形成具有倾向性的群体情绪，进而对股票市场价格产生影响。

其次，投资者情绪也受到股票市场的影响。投资者决策时其心理因素会随着股票市场的变化而改变。例如，当股票市场上充满着许多不确定性的时候，投资者会规避风险，试图进行理性的投资，然而投资者会发现自己的对股票市场的认知能力有限，为了进行更好的投资决策投资者会借助于媒体信息、专家建议以及自身的感觉、经验等。股票市场的不确定性越大，投资者的这种求助感越强烈，人类的认知偏差就越可能出现，从而导致投资者的非理性行为。

由此可见，投资者情绪与股票市场价格是相互改变、相互影响的。

2. 获取投资者情绪分析的方法

应用网络舆情分析投资者情绪，需要从大量文本信息或非结构化数据中挖掘有价值的资料。通过网络舆情分析投资者情绪的过程如图 4.9 所示。

首先，应用文本挖掘技术，从杂乱无序的网络媒体信息中获取有价值的信息，把非结

构化的文本信息转化为结构化文本信息，从文本信息中提取投资者情绪测评指标，结合属性词典和情感词典，应用情感分析引擎，获得投资者情绪分析结果。然后，可支撑两方面的应用：一是基于投资者情绪分析结果，以及情绪与股票市场之间走势的关联。对市场行情进行预测。二是基于文本信息中的上市公司属性和投资者情感倾向，预测各类上市公司的股票价格走势，为买入、观望、卖出等决策提供支撑。

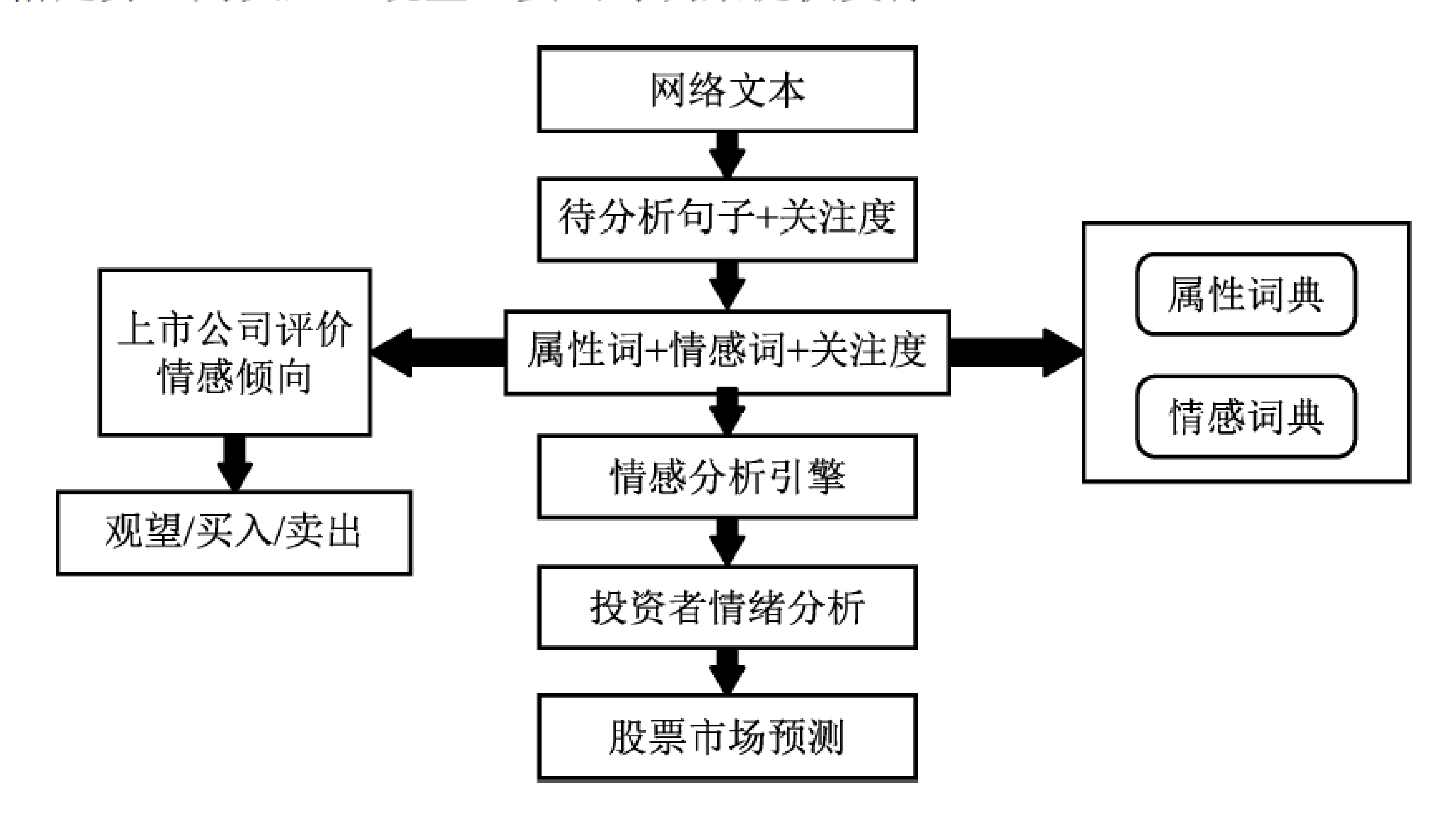


图 4.9 通过网络舆情分析投资者情绪

对于网络舆情中投资者情绪的分析，主要应用网页抓取技术、特征挖掘技术以及情感极性分类技术等。

1) 网页抓取技术

网络爬虫是目前使用最多的文本采集技术。网络爬虫又称为网络蜘蛛，是一个自动抓取网页的计算机程序，作为搜索引擎的重要组成部分来使用，为搜索引擎从互联网下载网页。通用网络爬虫的原理如下：从一个或若干初始网页的 URL 开始，获得初始网页上的 URL 列表，在抓取过程中，不断地从当前页面上抽取新的 URL 放入队列，直到 URL 的队列为空或满足某个爬行终止条件。主体爬虫的工作流程较通用网络爬虫复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的 URL 队列中。然后，根据一定的搜索策略从队列中选择下一步抓取的网页 URL，并重复上述过程，直到满足系统设置的某一停止条件。有别于传统网络爬虫的是，主体爬虫主要解决 3 个问题：一是对抓取目标的描述或定义；二是对网页或数据结构的分析与过滤；三是确定对 URL 的搜索策略。这一过程所得到的分析结果还将对以后的抓取过程提供反馈和指导。

优秀的网络爬虫工具应当具备抓取速度快、抓取准确率高、更新及时、可拓展性强、具有分布式抓取等特点。目前比较流行的抓取工具包括：Heritrk、WebSphinx、MetaSeeke 等。

2) 特征挖掘技术

特征挖掘技术是一种能够从结构化的文本信息中提取出关键属性词的技术。属性词一般由名词和名词短语组成。例如，“贵州茅台(600519)关于部分监事辞职的公告”其中“部分监事辞职”就是一个投资者关注的属性词。产品具有多种属性，也称为产品特征。

一般情况下，一篇产品评论信息可能涉及产品的多个特征。相应地，上市公司也具有不同属性，在股吧评论信息中涉及上市公司的不同的属性。例如：产品、业绩、利润等。

产品特征可以分为显性特征和隐性特征两类。显性特征是指出现在语句中可以直接作为产品特征的词汇或短语，而隐性特征是指句子中没有明显的特征描述，需要对句子进行语义理解后才能得到的特征。提取隐性特征需要自然语言的完全理解技术，而目前该技术还不够成熟。因此，目前的产品特征挖掘只考虑显性特征。在网络舆情中也只能识别上市公司的显性属性，进而判断投资者对不同显性属性的情感倾向。

特征挖掘技术是技术框架中的重要内容，目前主要有两种技术方法。第一种是人工定义的方法，这也是最常用的方法，主要有以下几种模式。

(1) 先应用文本特征表示，再建立挖掘模型，类似于文本关键词的提取方法。

(2) 先建立概念模型，再根据评论信息中的语音进行模式匹配。

(3) 建立领域知识模型，比如某些研究中挖掘出抽象属性就是应用了事先建立的领域知识模型。

(4) 建立本体模型，这一类研究和概念模型比较接近，就是事先建立了一个关于产品的相关概念及关系的本体。目前在这一领域应用得比较多的英文词网就相当于一个通用的语言本体。

第二种是自动提取的方法。该类方法主要通过词性标注、句法分析、文本模式等自然语言处理技术对评论信息进行文本分析，自动发现文本特征，这种方法具有很强的可移植性。

从挖掘效果上看，自动提取办法的结果通常查全率比较理想，但是查准率与人工定义法仍有一定差距。

3) 情感极性分类技术

情感极性分类主要是分析主观性文本、句子或者短语的褒义或贬义，即判定它们的极性类别。情感极性分类是有指导的机器自动分类，一般分为训练和分类两个阶段，可以分为以下几个步骤。

(1) 确定情感分析单元。情感分析单元即情感极性的分类对象，它是由研究目的所决定的。情感分析单元选择是否合适，直接对文本信息的情感分析效果产生较大的影响。

情绪单元可以分为词汇短语层、句子层和文档层 3 个粒度层面。

① 词汇短语层。它主要研究集中在单个词语或短语的语义倾向性，采用的方法主要包括基于语料挖掘的方法和基于极性词典拓展的方法。

② 句子层。情感可以由主题、意见持有者、情感描述项和褒贬倾向性的 4 个部分来描述，即意见持有者针对主题表达了具有某种褒贬倾向的情感描述。语句的情感分析重点是在语句文本中自动确定这些元素以及它们之间的关系的过程。

③ 文档层。文档层情感分析一般首先计算或判断词汇或词组的褒贬倾向性，再通过篇章中极性词语或词组技术或对其褒贬程度值求和或均值或结合句法分析等获得句子或篇章的总体情感极性。

(2) 文本表示训练文本。文本表示将决定选用什么样的文本特征来表达文本信息。就

目前的文本分类系统来看，绝大多数都是以词语或者词语组合作为特征项表达文本信息的。

(3) 挑选分类方法并训练分类模型。已有的文本分类方法可以分为：统计方法、机器学习方法等。在对待分类样本进行分类前，需要确定分类方法，利用训练文本进行学习训练并获得分类模型。

(4) 运用分类模型对测试集进行极性分类，评价所建立的分类模型的分类效果。

情感极性分类算法可以分为两类，即基于语义的情感分类方法和基于机器学习的情感分类方法。

① 基于语义的情感分类。是指通过文本信息语义分析的方式建立情感分类器，主要有两种方式。第一种是先从情感单元中抽取带有情感倾向的形容词或者动词，将其称为情感词，以及和这些词具有修辞关系的程度副词或否定副词，然后对这些情感词进行情感倾向计算，并得到它们的情感倾向值，最后对情感词的情感倾向值求和，得到情感分析单元的情感倾向值。第二种是建立一个包含情感字典的情感倾向语义模式库，然后把情感倾向分析单元按照这个模式进行模式匹配，计算出情感倾向值，最后对这些短语模式的情感倾向值求和，得到该情感分析单元的情感倾向值。

② 基于机器学习的情感分类。主要算法包括：朴素贝叶斯算法、决策树、人工神经网络、K 近邻算法等。对常用文本分类算法分析比较发现，支持向量机、K 近邻算法、朴素贝叶斯是 3 种较好的文本分类算法，其中支持向量机具有最高的分类精度，但分类速度最慢，朴素贝叶斯算法具有最高的分类速度但是精度最低。

基于语义的情感分类算法和基于机器学习的情感分类算法各有利弊。基于语义的极性分类算法能够更加接近现实的语义特征，但分析效果依赖于对语义模式的正确归纳；基于机器学习的情感分类算法，直接明确提取文本信息情感特征项，但分析效果依赖于语料库或训练文本信息的代表程度。

(5) 使用获得的分类模型对待分类文本进行分类，并对分类效果进行评价。

文本分类中普遍使用的性能评估指标包括查准率(Precision)和查全率(Recall)。查准率反映了一个分类器对于类别的区分能力，查准率越高，表明分类器识别的正确分类数与总分类数差距不大，即识别的错误率较低。查全率反映了一个分类器的泛化能力，查全率越高，说明这个分类器能够把正确的类别识别出来，但并不关心识别出的总个数。

为了判断属性词所在文本信息的情感极性是否符合人工标注的真实极性，可以归结为一个二值分类，评估选择使用二维列联表。判断情感极性的过程中可以通过列联表进行展示，如表 4.7 所示。真正属于该类的极性数即在人工标注中得到的情感极数。衡量查准率与查全率的计算方法如下：

$$\text{Precision} = \frac{A}{A+B}$$

$$\text{Recall} = \frac{A}{A+C}$$

表 4.7 评估极性分类性能的列联表

	情感极性句子数	非情感极性句子数
挖掘出来的情感极性句子数	A	B
未挖掘出来的情感极性句子数	C	D

如果算法的查准率高而查全率低的话，虽然分类效果的可靠性高，但对新的语句进行分类时很多正确的类别不能识别。而如果算法的查全率高查准率低的话，虽然对新语句的正确识别效果很好，但分类结果中错误的数量可能会比较多。由此分析，单独使用查准率和查全率中的一个指标来评价分类算法是不全面的，需要综合考虑。

4.4 大数据与量化投资

4.4.1 量化投资概述

量化投资(Quantitative Investment)，是指通过对金融市场和产品信息进行量化分析，根据历史交易和相关数据建立模型，由模型做出投资决定，再根据算法自动下单完成交易。

与其相对应的一个概念是定性投资(Traditional Investment)，它是指通过研究市场和金融产品信息，参考历史和当前该产品的交易价格，根据主观经验做出投资决定，进行下单交易。

量化投资和定性投资一样，也需要做交易前分析、下单交易和交易后分析等 3 个方面的工作。其中的人工工作包括建立数学模型、挖掘数据模式、开发计算机软件系统、设置各种参数，在量化投资软件系统运行后，还要对系统进行分析评估，然后根据评估结果调整模型或者重新挖掘数据模式，使得系统更加有效。

当下在金融领域出现多种灵活多变的量化投资策略，如量化选股、量化择时、量化套利、算法交易、资产配置等。

1) 量化选股

量化选股是指通过数量分析判断是否应该购入某种股票。具体的方法主要包括公司估值法、趋势法和资金法。①公司估值法是通过分析公司的基本面得出公司股票的理论价格，并通过与市场价格做比较从而确定投资策略。②趋势法是把市场分为强市、弱势、盘整 3 种形态，投资者根据不同的形态做出相应的投资决策。③资金法是根据市场主力资金的流动方向进行投资决策。

2) 量化择时

量化择时是指根据数量化的方法，对经济基本面进行量化分析的基础上，参考历史以及当前的市场价格，确定某只股票合适的买入时机。具体方法有趋势择时、市场情绪择时、牛熊线、Hurst 指数等。

3) 量化套利

量化套利是指运用量化分析的方法确定某种标的的最优投资组合，并将一种投资组合看成一种金融产品进行量化研究。主要包括股指期货套利、商品期货套利、统计套利、期

权套利等。

4) 算法交易

算法交易又称自动交易、程序交易或者机器交易，它指的是通过计算机程序发出交易指令。在交易中，程序可以决定的范围包括交易的时间、交易的价格等。

5) 资产配置

资产配置是指资产类别选择、投资组合中各类资产的适当配置以及对这些混合资产进行实时管理。量化投资管理将传统投资组合理论与量化分析技术相结合，极大地丰富了资产配置的内涵，形成了现代资产配置理论的基本框架。

量化投资的优势有以下几点。

(1) 大数据量的市场分析。这是投资决策的基础，定性交易靠的是人工调研，所以没有办法分析市场的所有产品。但量化投资可以分析市场的所有数据，从而可以获得更准确的市场信息，使得交易决策更科学、更系统、更有效。

(2) 快速交易。量化技术中引人注目的是快速交易，包括算法交易、高频交易。例如，在秒级时间内完成多个金融产品组合的下单交易、一分钟完成几个交易周期等，这些是手工方式根本无法想象的。更多的交易机会意味着更好的概率显著性，从而获得更好的投资收益。

(3) 理性交易。由于交易决策是由计算机程序做出的，不为人的主观情绪所左右，所以下单交易表现出良好的理性，好处是可以克服人性的弱点，如贪婪、恐惧、侥幸心理，使得投资更加理性。

4.4.2 证券量化投资中的主要分析工具

在金融领域中，量化投资的主要分析工具有数据挖掘、人工智能、小波分析、随机过程、分形理论、支持向量机等。下面介绍几个主要的分析工具。

1) 数据挖掘

数据挖掘是从数据库中获取信息的一个基本方法，其常用的方法有决策树、人工神经网络、关联分析等。模型也可分为聚类模型、关联模型、顺序模型等。数据挖掘常常应用于板块轮动策略中。板块轮动，指的是板块与板块之间出现轮动，推动大盘逐步上扬。比如，前一段时间金融板块率领大盘上涨，现在是地产板块推动大盘上涨，这就叫作金融板块与地产板块出现了板块轮动效应。由于股票市场经常出现板块轮动、涨跌不一的情况，因此可以利用基于关联规则的板块轮动策略进行投资。

2) 人工智能

人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能是对人的意识、思维的信息过程的模拟。它包括了机器学习、自动推理、人工神经算法、遗传算法等。在金融投资领域中，主要运用于短线投资。例如，同花顺软件的“智能选股”功能，就是基于人工智能的技术，推送投资者理论上有投资价值的股票。

3) 小波分析

小波分析其实是应用数学和工程学科中的一个概念，“小波”就是小的波形。所谓“小”是指它具有衰减性；而称之为“波”则是指它的波动性，其振幅正负相间的震荡形式。它能根据频率的变化调整分析窗口的大小。由于金融时间序列具有非平稳性、非线性的特点，因此传统的去噪方法效果不好，但小波分析可以克服这些缺陷。

4) 随机过程

随机过程是指选取一定的随机变量，通过观察表面的偶然性描述出必然的内在规律并以概率的形式来描述这些规律。研究随机过程的方法多种多样，主要可以分为两大类：一类是概率方法，其中用到轨道性质、停时和随机微分方程等；另一类是分析的方法，其中用到测度论、微分方程、半群理论、函数堆和希尔伯特空间等。研究的主要内容有多指标随机过程、无穷质点与马尔可夫过程、概率与位势及各种特殊过程的专题讨论等。对股市的大盘进行预测时，经常会用到马尔可夫链。

4.4.3 大数据在证券量化投资中的应用

大数据技术在证券量化投资中的应用可根据数据结构的不同分为结构化数据的应用与非结构化数据的应用两种应用方式。其中结构化数据的应用最为普遍。

1. 结构化数据的应用

在量化投资中，结构化数据应用主要集中于高频交易应用。高频交易(High Frequency Trading)是一种交易策略和技术，它是指从那些人们无法利用的极为短暂的市场变化中寻求获利的计算机化交易。例如，某种证券买入价和卖出价差价的微小变化，或者某只股票在不同交易所之间的微小价差。高频交易具有交易量大、交易次数多、持仓日短等特点，因此计算机每秒需要处理大量的结构化数据。此外，高频交易具有每笔收益率很低，但是总体收益稳定的特点，因此深受国际大型投资机构的青睐。

一般来说，高频交易可以分为两大类。

1) 传统的低频交易高速化

包括高频统计套利、高频阿尔法套利、高级趋势追踪等。其中高频阿尔法套利中的配对交易最为典型。配对交易是指从市场中寻找历史股价走势相近的股票作为配对股票，当股票的价格差偏离于历史均值时，则卖出其中股价较高的股票，买入股价较低的股票；当二者的价差回归历史均值水平时，分别平仓完成套利交易。另外，设置适当的止损点结束头寸以控制风险。配对交易具有广泛的应用性，除了股票这一标的资产外还可以应用到期货、期权、外汇等。

在配对交易过程中，获取大数据和大数据分析方法至关重要。首先，我们应该从市场中获取海量的交易数据，通过相关性分析方法找到价格相关走势高的证券；然后，根据海量的高频交易数据计算证券间的价格差，形成价格差的概率分布；之后依据概率分布设定触发条件和终止条件的阈值。例如，当证券价格差超过 X 临界值时开始买入卖出证券，当价格差到 Y 临界值时平仓。最后，根据设定，若某证券价格差持续扩大到 Z 止损点，可以选择平仓并止损。

2) 高频交易策略是凭借海量数据、高速交易而开发的新策略

这类策略的持仓时间非常短。例如，自动做市商策略利用量化算法优化头寸的报价和执行，其持仓时间只有 1 分钟。市场微观结构交易策略对观测到的报价进行逆向工程解析以获得买卖双方下单流的信息，该策略的持仓时间仅为 10 分钟。事件交易策略通过宏观事件进行短期交易，该策略持仓时间一般不会超过 1 小时。由此可见，高频交易一般不涉及隔夜持仓，因此它避免了隔夜风险。这在流动性紧张、隔夜拆借利率高的情况下更具有吸引力。而且基于计算机的决策算法与执行算法的结合能够有效避免人工决策时的情绪影响，这对提高整体的投资收益极为关键。更重要的是，高频交易策略拓展了投资的深度与广度，不仅充分挖掘了市场的潜在信息，而且拓展了市场范围。只要交易模型设计合理，就能在传统分析师不熟悉的市场上获得稳定的收益。

另外，开发高频交易策略也为投资者带来了巨大的挑战。首先，高频交易不仅数据量异常庞大，而且数据之间的时间间隔也不一致。传统的量化分析的方法完全不适用。其次，高频交易要求极高的准确性，交易信号如果延迟或者提前，投资者很可能在一瞬间由盈利转为亏损。最后，执行的速度是高频交易的核心。提高交易速度是各投资机构一致追求的目标，而更快的速度需要更大的资金投入。

可以看出，高频交易是未来证券投资领域的重要发展方向之一，其稳定的投资收益与科学的决策过程吸引了越来越多的投资者加入。目前，国外顶级投资机构 60%以上的交易都是通过高频交易完成的，并且这一比例还在不断扩大。在中国，随着金融市场管制进一步宽松，适合进行高频交易的投资品种正逐步增加，高频交易将会得到更多国内机构投资者的青睐。

2. 非结构化数据的应用

目前，非结构化数据在量化投资领域的应用并不普遍，但业界正在进行大量的尝试。非结构化数据能够提供有价值的信息并进而获得超额利润，这推动了更多的公司在这方面加大投入，并取得了一定的成果。

【案例 4.1】 伏流投资：掘金大数据，筑建量化投资

“我们先是一家数据科技公司，然后才是资产管理公司。”伏流投资强调，公司将大数据、量化分析和交易技术作为公司的基石——数据分析和数据科技是基础；量化模型分析建立在厚实的基础数据之上，致力于 SmartBeta、Alpha 的发现和获取；交易技术则为实现交易实践提供技术支撑，三者融合，缺一不可。

伏流投资目前拥有成熟的模型和策略，囊括不同周期、不同品种，可容纳资金量约 5 亿元。在公司内部，一套交易策略从研发到成熟，要经过严格的回测分析、黑箱测试、参数检验、失效检验、边界分析，最后进入策略库。他们认为，策略要保持一致性，符合公司收益风险特征；要观察策略在正常情况下 sharp 比率、换手率等指标的表现；评判策略是否失效则要通过市场检验。伏流投资以月为周期更新策略，调整策略参数，一旦失效则停止该策略。

伏流投资的量化模型建立在基本面和技术面数据之上，同时参考大数据舆情面数据指标作为参考，即考察市场参与主体和大众的情绪认知。此外，成交量也会作为量化模型的

参考。

“量化投资最大的风险来源于参数调优过度拟合、历史回测与实际交易偏差、量化模型失效以及极端行情风险。”伏流投资提到，针对这些风险，公司建立了事前、事中和事后完整的风控体系——事前规划、事中全程监控、事后复盘分析。

大资管时代，伏流投资亦走上自主发行产品之路，目前首只产品已经完成备案，该产品使用混合型策略，投资标的包括股票和商品期货。按照公司规划，伏流投资未来产品线将逐渐覆盖量化选股、量化对冲，海外产品、固定收益类产品亦在计划之中。

伏流投资认为，量化投资因为其风险收益可度量、可回测以及客观等特性，加之交易工具的进步，量化投资将成为主流趋势。在百舸争流的局面下，私募机构唯有在人才、策略、产品、营销、合作等各方面有综合优势，才能长远发展。

未来 3～5 年，伏流投资将着力多策略、全品种、全天候的研发，把人才培养、技术更新和产品绩效作为重点工作来部署，进一步拓宽市场的广度和深度，丰富投资策略，不盲目追求资产规模，稳健发展，为将来资产管理规模的扩展做好准备。

(资料来源：《期货日报》第 004 版,2016-07-20)

【案例 4.2】 机构选股逻辑基因变异　量化投资互联网掘金大数据

利用互联网大数据挖掘股市的超额收益机会正成为近期基金业的一股新潮流。而动作较快的当属广发基金和南方基金这两家基金公司。

其中，广发基金联合百度公司、中证指数公司开发百发 100 指数，南方基金则携手新浪财经、深证信息公司推出了 i100 指数和 i300 指数。

基金公司竞相开发大数据指数的动力在于，基于大数据筛选出来的组合，大幅跑赢现有的指数基金。

中证指数公司提供的数据显示，自 2009 年至 2014 年 6 月 30 日，百发 100 指数的累计收益率达到 545%。同期，中证 500、中证全指、沪深 300 指数的收益率分别为 102%、56%、19%。

历史收益源自模型样本的模拟测算，外界对其收益率或许存有疑问。而百发 100 指数产品在模型样本外的实盘数据，同样大幅跑赢主流指数。

2014 年 6 月 20 日开始，百发 100 指数进入实盘运行阶段。自此至同年 10 月 8 日，百发 100 指数实现的累计收益率达到 43.33%。同期，沪深 300 指数、上证指数和创业板全指的收益率分别只有 16.52%、17.74%和 17.94%。

i100 等权重指数和 i300 等权重指数的历史收益率，同样凸显出大数据的优势。

自 2010 年 1 月 29 日至 2014 年 7 月 31 日，i100 和 i300 的累计收益率分别达到 222.40%和 141.58%，远高于同期创业板指数 34.45%的累计收益率，更高于中小板指数 −8.95%的累计收益率。南方新浪大数据指数自 8 月开始正式进入模型样本外运行。其中，i100 等权重指数在 8 月和 9 月实现的月收益率分别为 10.93%、15.63%。这意味着该指数在两个月期间的收益率达到 26.56%。

券商的研究团队同样在挖掘“大数据”带来的投资机会。其中，长江证券金融工程团队自年初即建立新闻选股模型，自 1 月 12 日开始样本外跟踪。

长江证券金融工程主管范辛亭发布的研究报告显示，2014 年 1 月 22 日至 8 月 15 日，新闻选股模型累计的绝对收益率达到 52.45%，超越沪深 300 指数的 39.44%，超越中证 500 指数的 30%。

招商证券金融工程高级分析师夏潇阳利用深交所互动易披露的调研信息，构建中小板创业板调研组合。实盘跟踪的结果显示，自 2014 年年初至 9 月 30 日，该调研组合实现的累计收益率为 29.31%，跑赢中小板指数 14.91%，超越创业板指数的幅度是 8.53%。

无论是长江证券、招商证券构建的选股模型，还是广发基金、南方基金推出的大数据指数，其共同点在于引入网民对个股的搜索大数据作为选股因子。

当基金公司和券商研究将互联网金融的大数据作为选股因子引入模型，代表着资产管理机构在指数投资上重构选股逻辑。统指数编制依赖的是市值规模、成交金额、财务及估值等传统因子。它最大的缺点是采用过去 3 个月或 6 个月的数据去预测未来一期的收益，参数对历史数据有严重的依赖。

利用百度数据融入了投资者在投资决策前的行为规律，对未来的市场投资规律有一定的预测作用，其预测效果好于传统的来源于历史数据的因子数据。与传统指数不同的是，百发 100 指数的编制思路是跳出行业、板块的限制，从全市场中寻找超额收益的机会。其选股模型的特点是，它所挑的股票是契合未来市场或行业轮动热点，且基本面良好，未来有一定成长空间的价值型股票。不过引入大数据的模型能否经受考验，还有待时间验证。

(资料来源：《21 世纪经济报道》第 023 版，2014-10-10)

本章总结

- 大数据技术已经在证券行业中得到了广泛的应用，主要应用于在股票分析、客户关系管理、投资情绪以及量化投资四个方面。
- 在股票分析中大数据技术是进行基本分析和技术分析良好的工具，主要运用的是数据挖掘的方法，例如决策树法、聚类分析法、人工神经网络算法、时间序列分析以及关联分析等。
- 在证券客户关系管理中，通过大数据技术可以构建客户细分模型(DFM 模型)将客户进行合理的分类，以便有效地对客户进行管理。证券公司一般以客户证券账户资产以及交易活跃度作为评定客户等级的主要标准，不同等级的客户其服务策略不同。

  此外证券公司可以构建客户满意度模型来分析存量客户对公司的满意程度，构建客户流失预测模型(以 Logit 为方法)了解客户流失的情况以及导致客户流失的原因，推动公司形成有效的决策，提高服务质量。
- 大数据技术也可以应用于衡量投资者的投资情绪。证券公司可以通过一些量化的主观情绪测量指标以及客观情绪测量指标了解投资者的投资情绪，运用应用网页抓取技术、特征挖掘技术以及情感极性分类技术等方式在网络舆情中获取重要信息。从而为其自营投资业务提供有效参考。

- 大数据技术在证券行业中最为广泛也最为重要的应用就是量化投资，大数据技术为证券投资提供多种投资策略，例如量化选股、量化择时、量化套利、算法交易以及资产配置等。通过数据挖掘、人工智能、小波分析、随机过程、分形理论、支持向量机等分析工具，使让证券投资实现了高频化、智能化。

本章作业

1. 在股票基本分析中，主要的分析因素有哪些？
2. 试述大数据在股票基本面分析和股票技术分析中的应用都有哪些方法，并进行简要的介绍。
3. 谈谈什么是客户细分，并简要介绍大数据技术在客户细分中的应用。
4. 证券公司客户流失的原因是什么？简要介绍流失客户模型建立的过程。
5. 什么是网络舆情？网络舆情与投资者情绪之间有着什么样的关系？
6. 试介绍国内外常见的投资者情绪指标并说明分析网络舆情中投资者情绪的流程。
7. 什么是量化投资？量化投资都包括哪些策略？
8. 在证券行业中，量化投资是怎么样实现的？试述量化投资的优势。

第 5 章

大数据在保险业中的应用

本章目标

- 掌握大数据保险的特征、应用阶段和主要作用
- 掌握大数据在保险承保定价中的应用
- 掌握大数据在保险精准营销中的应用
- 掌握大数据在保险欺诈识别中的应用

本章简介

随着大数据时代的到来，大数据技术逐渐渗透于各个行业之中，并不断地颠覆传统的行业管理和运营思维。作为大数据的生产者和使用者，保险行业也在积极应用大数据技术，但保险行业的大数据应用才刚刚起步，与银行业和证券业相比其应用大数据的能力还相对落后。这主要是因为保险行业的数据基础尚未完善，其内部数据大多仍处于数据孤岛的状态，致使其内部数据难以被充分挖掘和使用。目前，大数据技术在保险行业中的应用主要体现于合理的承保定价、精准的保险营销和有效的欺诈识别。本章将重点讲解大数据在保险行业的承保定价、精准营销和欺诈识别中的作用。

5.1 大数据保险

5.1.1 大数据保险的概念和特征

保险的业务特点使其天然就具有大数据的特征，具体表现在以下 3 个方面。

(1) 保险业是经营风险的行业。

由于保险业是经营风险的行业，因而其所经营的保险产品在设计时需要对标的物的风险进行精准测定。而风险测定要以充分的数据为基础，保险公司自身已掌握海量的数据，需要利用大数据技术对这些海量数据进行分析从而有效地量化风险。

(2) 对未来风险的预测是保险公司的利润来源。

保险公司的利润来源于其向投保人所收取的保费与相应标的物未来发生的赔付支出之间的差额，因而保险公司需要对相应标的物未来风险发生的概率进行预测。而预测正是大数据的核心功能，与保险经营的关键需求不谋而合。

(3) 保险经营的过程中包含着数据的产生与使用。

保险经营的过程包括产品设计、产品营销、承保定价、风险防控、核保理赔等一系列环节。在这些环节的具体运行过程中，大量的相关数据被不断地利用，更多新的可利用数据也在这一过程中不断产生。

1. 大数据保险的概念

大数据保险是指保险公司通过利用大数据技术对风险数据进行分析、处理和挖掘，使风险数据实现有效的价值变现。在此基础上保险公司通过其治理端和商业端的协同创新，使传统的保险服务方式和资源配置方式得以优化，从而实现保险产品、保险服务和保险业务模式的创新，进而更好地满足其客户需求并提供更为优质的保险服务。

2. 大数据保险的特征

大数据保险所具有的特征表现为以下 6 个方面。

1) 数据驱动

与互联网保险的渠道驱动所不同的是，大数据保险是由数据驱动的。保险数据处理技术的变革和应用是大数据保险发展的关键驱动力。大数据技术不仅可以在保险公司建立风险模型和对产品进行定价的过程中被充分利用，也能够在承保理赔过程中的各环节发挥作用。

2) 问题思维

在运用大数据技术实现数据挖掘和数据价值变现的过程中，大数据技术消灭信息不对称、不匹配的能力得以体现。保险公司在业务开展过程中所遇到的难点和痛点，正是应用大数据技术的重点；通过利用大数据技术对数据进行分析和处理，之前的难点和痛点将变为大数据保险的创新点。

3) 融合创新

大数据技术在保险领域中的应用使保险业在与新技术相融合的过程中，推出更多具有

创新性的产品和服务，也使保险公司的业务模式得到了创新和优化。

4) 运营提升

通过利用大数据技术，大数据保险的资金摩擦被最小化，资源配置的过程得到充分的优化，进而使其运营效能得到了有效的提升。

5) 活力生态

随着大数据技术与保险行业的深度融合，数字生态系统的建立势在必行。在这一生态系统中不仅有保险公司的参与，还会有其他行业的参与者参与其中。数据在这一生态系统中不断地更新，从而使得该生态系统更加具有活力。

6) 服务导向

在传统保险中，虽然众多保险公司早已将针对客户需求的服务导向作为其经营的核心价值观，但碍于时间与空间上的信息不对称，该服务导向在重重制约中被扭曲。大数据保险通过利用大数据技术在交互的价值网络中及时有效地获取信息，实现了信息数据的透明化，进而帮助保险公司提供真正从客户需求出发的保险服务。

5.1.2 保险业大数据应用的阶段

1. 世界保险业的数据应用发展阶段

保险业的经营和发展与数据应用密不可分，因而保险业数据应用沿革的历史也是其发展的历史。根据世界保险业在不同时代数据能力和技术水平的不同，我们可以将其数据应用发展过程分为 4 个阶段，分别是数据匮乏时期、统计数据应用时期、信息技术应用时期和大数据技术应用时期，如图 5.1 所示。

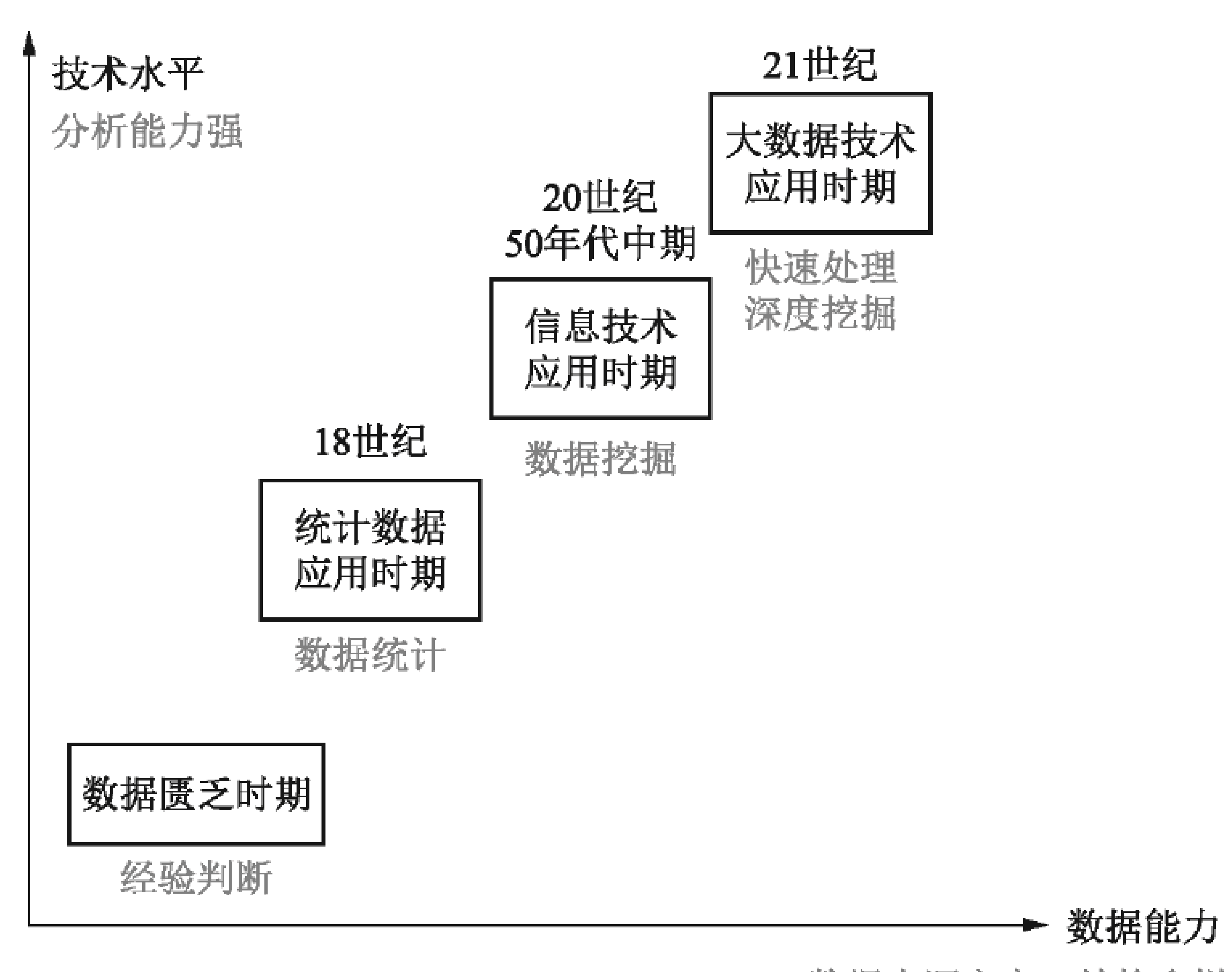


图 5.1 保险业数据应用的发展阶段

(1) 18 世纪之前，世界保险业处于数据匮乏时期。当时承保人获取信息的渠道非常有限且信息可靠性较低，风险定价主要依赖于承保人的经验判断。

(2) 18 世纪寿险生命表和均衡保费理论出现，世界保险业进入统计数据应用时期。在这一阶段数学方法和统计手段开始应用于保险定价中，从而使寿险业得到快速发展。

(3) 随着信息技术在 20 世纪 50 年代中期的快速发展和广泛应用，世界保险业进入信息技术应用时期。在这一阶段保险经营过程中所依赖的数据基础得到不断的扩充，行业数据应用水平也在日益提高。

(4) 进入 21 世纪，随着移动互联网、社交网络、大数据等新技术的出现和快速发展，世界保险业数据应用进入了大数据应用时期。在这一阶段保险业所掌握的数据从内部数据扩展到外部数据，从定量数据扩展到定性数据，从结构化数据扩展到半结构化数据和非结构化数据，从交易数据扩展到行为数据，数据来源不断丰富，数据结构更加多样。

2. 保险业大数据应用的阶段

结合大数据技术的发展趋势，可以将保险业大数据的应用分为 3 个阶段：内部循环、外延拓展和全面应用，如图 5.2 所示。

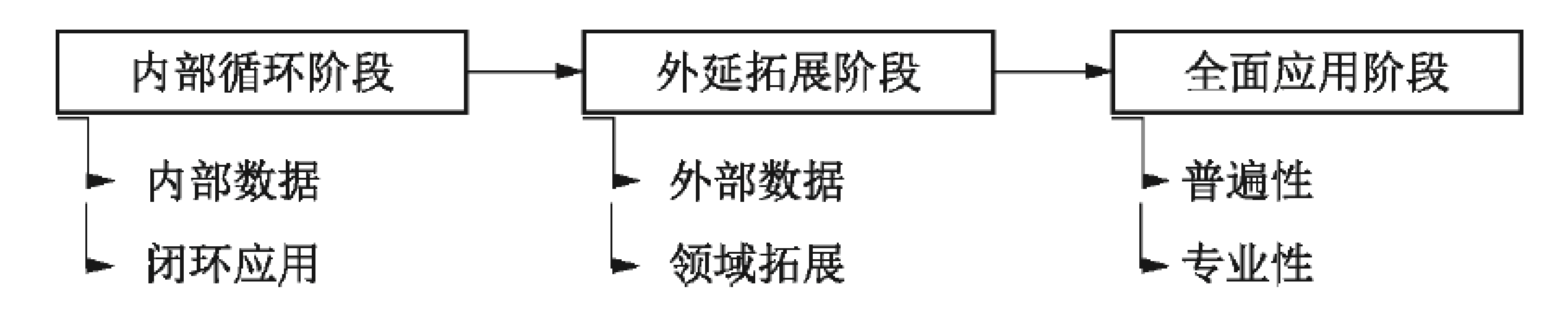


图 5.2　保险业大数据应用的阶段

(1) 内部循环阶段。保险公司利用其在业务经营活动中所产生的大量内部数据，通过利用大数据技术进行深度的挖掘分析，实现以数据指导决策，帮助业务流程有效优化。在此基础上更多的客户被吸引并带来更多新的可利用数据，从而形成具有正向激励特征的闭环。

(2) 外延拓展阶段。保险公司开始尝试利用内部数据解决其主要产品及服务以外的问题，进而拓展其内部数据的应用领域；或是引入与其主要产品和服务直接或间接相关的外部数据，通过利用大数据技术进行充分的挖掘和分析更好地解决其发展中所遇到的问题，并为其提供更多的创新机会。

(3) 全面应用阶段。经过行业相关数据的规模化和规范化发展，在行业数据产业链上分化出数据提供者、数据加工者、数据消费者等专业化组织。在这一阶段中，数据来源愈加丰富化，数据结构愈加多样化，大数据技术的应用也更加具有普及性和专业性，行业技术水平和分析能力也在不断提高。

5.1.3　大数据在保险行业中的作用

随着大数据技术与保险行业的逐渐深度融合，保险公司将实现对大规模、多样化的数据的及时获得和快速分析。在可预见的未来，保险产品和服务的性质将会发生根本性的变化，即保险价值将会更多地体现在后端专业化的风险解决方案上，而不再是风险条件触发

后的赔付。

1) 产品和服务的个性化

在传统的保险经营中，保险产品和服务的设计、营销、推广等环节仅关注于具有相似特征的某一客户群体，而不是具体的单一客户。保险公司通过利用自然语言识别、文本挖掘、模糊判断等大数据技术，可以对单一客户在社交平台上留下的海量数据进行挖掘和分析，从而了解该客户的行为习惯、风险偏好和保险态度，进而为其提供个性化的保险服务和精准的风险控制。

2) 风险衡量的精准化

保险公司借助大数据技术，可以实时地对与单一客户相关的海量数据进行挖掘和分析，从而实现对该客户实时且精细化的风险衡量。例如，对于车险客户，保险公司通过对其所获取的客户驾驶行为信息、车辆行驶信息和交管局的违章信息等信息数据进行处理，精准地衡量出该客户当日的风险状况进而计算出其当天应缴纳的保费。

3) 保险价值链的再创新

大数据技术的应用使保险公司的外部交易成本得到大幅降低，进而帮助保险公司实现资源的有效整合，促使保险价值链实现再创新。

4) 供应商的优化整合

保险公司在其长期的经营过程中积累了大量的客户数据。在借助大数据技术对这些海量客户数据进行整合和分析的基础上，保险公司可以与汽车修配企业、医院、药品生产企业等相关机构开展更加深度的合作，在降低其经营成本的同时为客户提供更加便捷的服务。

5) 保险需求的发现和引导

在当前快速发展的信息时代，人们在依托互联网所建立的社交平台上发布信息、交流观点和表达想法。保险公司通过借助大数据技术对这些社交平台上的信息数据进行挖掘和分析，能够及时有效地获取人们的关注点和行为偏好，进而找出潜在的新保险需求，设计出有针对性的保险产品和服务，实现对客户保险需求的及时发现和有效引导。

6) 商业机会的有效发掘

保险公司在数据方面具有得天独厚的优势。在当前的大数据时代，保险公司通过利用大数据技术对其所掌握的大量业务数据进行分析、挖掘，可以对其所获得的数据处理结果加以利用，从而发掘出更多的商业机会。例如，保险公司可以建立销售平台向消费者出售适当的车辆维修保养服务。

7) 企业生态系统的再构建

保险公司在利用大数据技术的过程中，其与外部市场之间的边界日趋模糊，即保险公司开始尝试与其他行业领域进行融合和合作，从而构建起基于大数据的企业生态系统。

8) 行业格局的快速变化

随着互联网与人们生活的日趋紧密，越来越多的互联网公司借助其在大数据利用方面的优势进军保险行业。例如，阿里集团、腾讯、京东等互联网公司均已开通其保险平台。此外，一些从事保险中介服务的机构也建立了线上的“保险超市”专门销售各类保险产品，它们依托大数据技术为其客户提供专业的保险业务咨询和个性化的保险方案定制服务。

5.1.4 大数据下的数据服务架构

1. 调整前后的数据服务架构对比

在大数据技术应用以前，保险公司为满足其业务经营的需要已针对其所掌握的结构化数据建立起一套数据服务架构。在当前的大数据技术应用的背景下，保险公司为实现对海量半结构化数据和非结构化数据的处理和分析，需要结合相关大数据技术对其数据服务架构进行调整。保险公司调整前后的数据服务架构对比如图 5.3 所示。

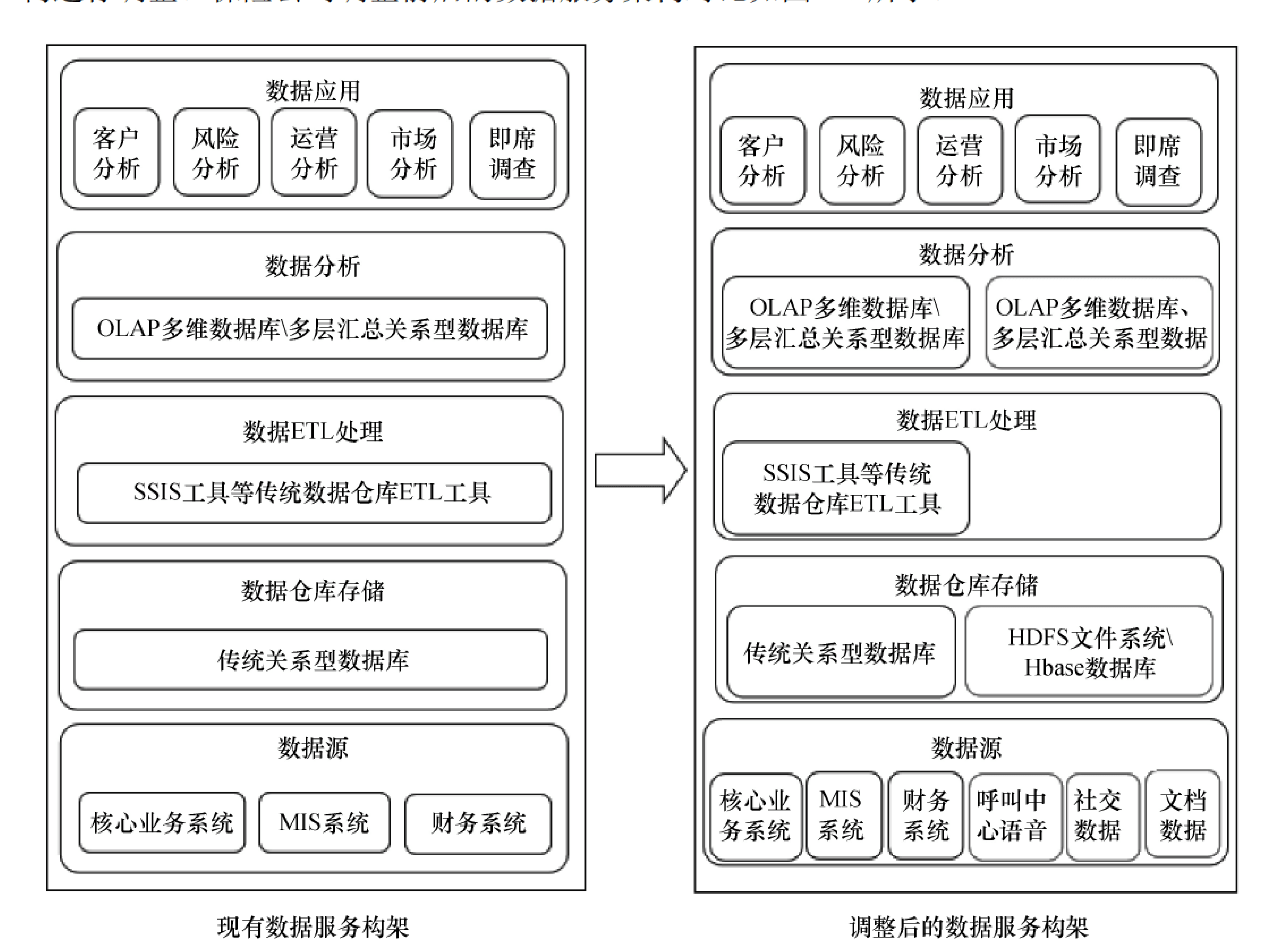


图 5.3 保险公司数据服务架构的前后对比

调整前后的数据服务架构之间的主要区别在于 Hadoop 技术被引入数据服务架构：在数据存储方面，增加了 HDFS 文件系统和 Hbase 数据库；在数据分析方面，增加了 MapReduce、Hive、Pig 等技术对存储数据进行计算和分析。

调整后的数据服务架构的主要变化具体表现为以下 3 个方面。

1) 数据源范围扩大

在数据源中增加了半结构化数据和非结构化数据，呼叫中心记录的语音数据、客户社交数据、相关文档数据等都被纳入其中，使保险公司各类数据的商业价值被最大化地利用起来。

2)　数据存储工具增加

增加了 HDFS 文件系统对分布式文件数据进行存储管理，以及 Hbase 数据库对海量结构化数据进行存储管理，从而使保险公司的经营管理能力随数据量的不断增长而逐渐提升。

3)　数据分析工具增加

在数据分析过程中增加了 MapReduce、Hive、Pig 等技术，从而实现对 Hadoop 数据的分析和计算。

2. 调整后的数据服务架构

从图 5-3 中我们可以看到，调整后的数据服务架构将采用传统数据库技术与 Hadoop 技术相结合的方式来满足保险公司的数据处理需求。传统数据库技术与 Hadoop 技术的结合方式可视保险公司的实际需要灵活调整：既可以选择用传统技术处理结构化数据，用 Hadoop 技术处理半结构化和非结构化数据；也可以选择将传统数据库中的结构化数据导入 Hadoop 之中，进而借助 Hadoop 技术来提升保险公司对海量数据的处理能力。

5.1.5　保险业大数据应用现状

1. 总体特点

在金融领域中，保险行业应用大数据相对较晚，应用水平也落后于银行业和证券业。这是因为银行业与证券业的数据服务平台建设较早，从而为大数据技术的应用奠定了良好的基础，而保险业的数据服务平台建设则相对较晚。

而就保险业自身的大数据应用阶段而言，目前尚且处于大数据应用的初级阶段，即内部循环阶段。因而接下来保险业需要通过合理利用其内部数据并引入更多的外部数据来拓展大数据分析在本行业中的应用领域。

从全球范围上来看，国外保险业的大数据应用水平高于国内保险业。

2. 国内保险业大数据应用的特点

国内保险业的大数据应用具有 4 个特点(见图 5.4)，具体介绍如下。

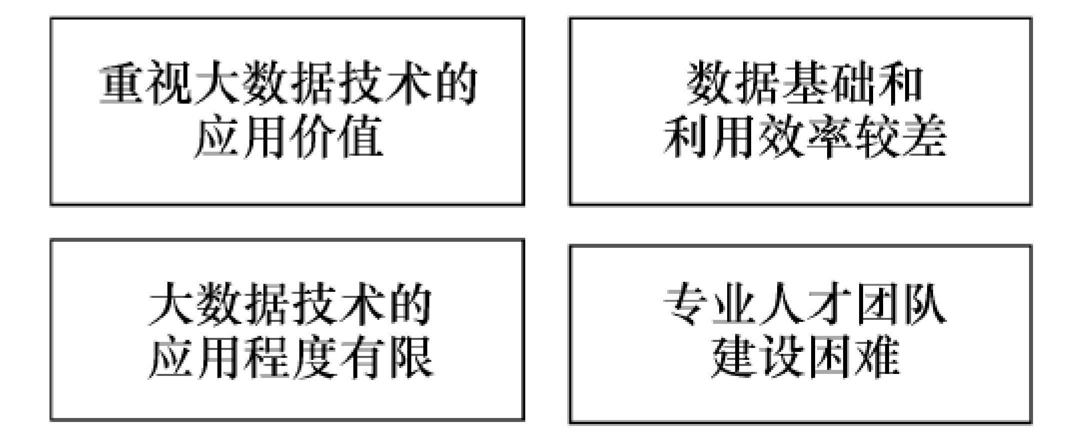


图 5.4　国内保险业大数据应用的特点

1)　重视大数据技术的应用价值

目前，国内保险业已经对大数据技术的重要应用价值形成广泛共识，认为大数据技术将给传统保险业带来深刻的变革，大数据技术的应用能力也将成为保险公司未来的核心竞争力。

2) 数据基础和利用效率较差

目前国内保险业的相关数据积累还十分有限。根据权威机构的调查显示，我国保险业的数据资源总量较少，且主要以结构化数据为主。虽然保险公司已积累一定量的半结构化数据和非结构化数据，但对这些数据的利用效率仍然较低。

3) 大数据技术的应用程度有限

目前国内大部分保险公司尚且处于对大数据技术的学习理解阶段，虽然已有少部分保险公司开始了对大数据技术应用的小规模试验，但尚未出现大规模的大数据商业应用。此外，国内保险业对大数据技术的应用主要集中在营销领域，应用范围也较为有限。

4) 专业人才团队建设困难

目前国内只用少部分的保险公司建立了专门从事大数据研发的团队。而且大多数保险公司的大数据研发人员主要来自信息技术部门，缺少同时具备金融保险知识和信息技术素养的跨学科复合型人才。

3. 保险业大数据挖掘所面临的主要问题

1) 数据的对接

由于金融行业十分重视数据的安全性，因此行业内的相关数据都具有较高的保密性。所以，如何与金融同业机构在确保相关数据安全的基础上进行数据的共享就成为保险公司在大数据挖掘中所要解决的问题。

2) 数据的考量

由于人们的行为活动在不断地进行，因而客户行为的相关数据也在不断地产生之中。保险公司会基于数据分析对客户进行画像，但很难通过几个固定的标签就描绘出客户画像，即客户的标签特征也在不断地变化之中。在不同的场景下，同一个客户可能分别被定义为有需求客户和无需求客户。因此，保险公司在大数据挖掘过程中要注意实时数据的识别问题。

3) 数据的应用

在做好数据的获取和挖掘工作后，如何有效地利用大数据技术应用下的数据分析结果成为保险公司所要回答的问题。目前，国内保险业的大数据应用主要集中在营销领域，因而有待大数据在产品设计、承保定价、核保理赔、风险防控等领域也发挥深层次的作用。

4. 保险业大数据应用的潜在突破口

1) 承保范围的扩大

在大数据技术的应用下，过去不可承保的风险也将有可能成为可以承保的风险，更多潜在的时新的保险需求也将被有效激发。目前已为人们所熟知的退货运费险正是大数据在保险业中应用的产物。

2) 个性定价的实现

在大数据应用的背景下，随着保险公司所掌握的数据在数量上日趋庞大、在维度上日趋宽广，其保险定价的精确度也日趋提高。这是因为保险公司通过应用大数据技术使其所面临的逆向选择风险得以降低，产品定价的优化和保险费率的个性化制定也将得以实现。

3) 核保理赔的优化

保险公司通过利用大数据技术对海量数据进行分析建模，可以使自动化的核保理赔过程得以有效实现。例如，北京市保监局与北京市交管局联合推出的 APP——“事故 e 理赔”就是大数据在保险核保理赔环节的典型应用。

4) 风险防控的提升

作为经营风险的企业，保险公司在其日常经营过程中面临着诸多风险。这些风险不仅表现为投保人的逆向选择问题和道德风险，也表现为保险公司自身的操作风险、信用风险等相关风险。保险公司通过有效运用大数据技术，可以使其风险管理能力和水平得到提高。例如，将大数据技术应用于核保理赔环节，能够有效地提高保险公司的欺诈监测能力。

5) 运营效率的提高

保险公司还可以将大数据技术应用到相关运营环节当中，如人力资源管理、财务管理等，从而实现保险公司运营管理水平的有效提升。

5. 案例：大数据与保险业务模式创新

通过运用大数据技术，国外保险公司在业务模式方面有了诸多创新，具体有以下几个方面。

1) 客户参与度的提高

澳大利亚的 Youi 保险公司为了提升其客户满意度，借助大数据技术对其客户反馈方式进行改进。Youi 保险公司将其客户评价实时公开在其网站上，客户在浏览其网站时不仅可以在屏幕下方看到关于该公司的最新评价，还可以通过选择关键字对所有评价进行过滤。Youi 公司通过引入该方法在客户服务方面取得了成功。

2) 车载信息系统的利用

Progressive 保险公司利用车联网推出了 UBI 车辆保险产品，从而大幅降低了其车险投保人的投保费率，并在此基础上为其投保人提供一系列与车辆相关的增值服务。

3) 保险生态系统的建立

中国平安保险集团依托大数据技术建立了其完整的保险生态系统，其所提供的保险产品和服务仅仅作为该生态系统中的一部分参与其中。在提供传统保险服务的基础上，该生态系统为了最大限度地挖掘客户潜力并保有客户，还能够提供创新服务。

4) 基于社交网络的互助保险

Friendsurance 公司和 RiskHuddle 公司基于社交网络向其客户提供互助保险，涉及个人责任险、家具险等险种。在该互助保险中，基于社交网络所形成的小团体，团体成员相互承保并形成资产池；当某一成员出险需要理赔时，相关款项先由该资金池支付，资金池不能覆盖的部分才由保险公司支付。在该模式中客户的投保费用和保险公司的理赔风险都得到了降低。

5) 销售方式的创新

Bought by Many 公司通过利用大数据技术来识别潜在客户的特定保险需求，并将其识别到的潜在客户需求与众多保险公司的保险产品相匹配，继而通过社交媒体和搜索引擎将所匹配的保险产品营销给该类潜在客户。在此销售方式下客户的投保开销和保险公司的客

户流失率都得到降低。

5.2 承保定价

近些年来，保险业所面临的外部环境发生了诸多变化，这些变化主要表现在人口结构、技术创新、金融产品和服务的融合发展等方面。为了应对这些变化，保险公司可以借助大数据技术来提高其获取和深度挖掘信息的能力，通过对客户的交易行为进行记录、分析和预测，提高其承保定价能力。

5.2.1 大数据与传统保险定价理论

1. 大数据与大数法则

保险作为一种风险管理的工具，是建立在社会群体之间的风险救助机制。而保险产品的设计机理主要是基于统计学范畴中的大数法则，即基于风险发生和损失的历史数据进行分析和预测，在重复地随机现象中找出具有一定必然性的规律，进而依靠精算技术对产品进行定价并建立合理的财务运行机制。大数据与大数法则虽然都是在大量数据基础上进行风险和财务预测，但二者在保险产品定价机制中的作用基点是完全不同的。

大数法则是保险定价的根本法则，特别是对于车险、寿险、健康险等关系社会公众利益的领域，保险公司必须依托大数法则来确保其行业基准纯风险损失率的厘定是公平、充足且安全的。即大数法则是保险运行管理的数理逻辑，是保险业不可动摇的理论和定价基础。而大数据则主要在保险定价中发挥辅助作用，特别是通过采集和获取客户交易行为、对相关网络数据进行关联分析，找寻数据背后风险与成本、收益的匹配规律，进而推动保险公司客户细分化、责任碎片化、产品定制化，优化精算定价模型，从而建立科学、有效的保险费率浮动机制和差别化定价机制。

因此，大数据并没有颠覆大数法则，而是对市场化保险费率形成机制的重要优化和改进，是一种以新技术为依托、更加精细化的风险管理辅助工具。

2. 大数据与传统保险精算理论

保险作为经营风险的学科，其运行的关键在于精算。在传统精算理论中，精算师通过运用大数法则对其所掌握的风险暴露数据进行建模和分析，从中找出该项风险发生的规律，并在一定的假设条件下对未来风险发生的可能性以及所造成的损失大小做出判断，进而基于这些判断设计相应的保险产品。

在大数据应用的背景下，精算师可以利用大数据分析技术对其所掌握的海量数据进行回归分析，进而精准地识别出具体某一客户的潜在风险，而不再是对具有相似特征的某一类型客户群的潜在风险进行判断。虽然二者在思维模式上有很大的不同，但大数据并没有颠覆传统精算理论，而是作为一项辅助工具与传统精算方法相融合，进而衍生出更加优化的保险精算方法。

5.2.2 大数据对承保定价的革新

1. 丰富风险特征的描述

在传统的保险定价方式中，精算师利用到的数据仅限于保险行业中的数据，甚至仅为保险公司的内部风险数据。在当前感知更加透彻、互联互通更加全面、智能化更加深入的大数据时代，大数据技术将帮助保险公司获取到丰富的风险特征描述，进而助其在承保定价方面实现革命性的创新。

1) 从样本数据到全量数据

保险精算是基于一定量的数据实现的。在传统的保险精算中，假设通过抽样所选取的样本是能够充分反映被调查群体特征的，但鉴于技术和操作层面所存在问题，基于样本的判断往往不尽如人意。而在当前的大数据时代，保险公司可以充分地利用依靠大数据技术所获取的全量数据，从而使保险精算更加准确。

2) 从内部数据到外部数据

一直以来，保险精算所利用的数据大多都是保险行业的内部数据，包括基于承保的风险数据和基于理赔的损失数据。传统保险精算就是在这些数据基础上进行分析建模从而对保险产品进行定价的；但就单独的风险个体来看，这些内部数据根本不足以刻画其个体风险。而在大数据技术的应用下，被引入的外部数据能够充分地丰富风险刻画的维度，并将会在保险公司的承保定价中发挥更加重要的作用。

3) 从历史数据到实时数据

在传统的保险精算中所利用的数据大多是历史数据，由于这些历史数据缺乏时效性，在其基础上所进行的保险精算并不能很好地满足预测和定价的需求。例如，我国的寿险业在过去一直使用的是日本 1965 年数据编制的生命表，显然与我国的实际情况存在较大的差距。而在大数据技术的应用下，保险公司可以实时地获取与保险经营相关的数据，从而实现更加精准的风险预测和定价。

4) 从数据数量、质量到维度

在大数据技术应用以前，人们在进行保险精算时通常都希望获取尽可能大的数据量并重视数据质量的把控工作。而在当前的大数据时代，数据的数量和质量不再是数据工作的关注焦点，因为大数据技术填补了过去数据维度有限的不足，使数据维度得到了极大程度的丰富。保险精算也将会把工作重点转移到利用多维度数据更好地刻画客户中来。

5) 从因果关系到相对关系

传统的保险精算是基于因果关系对历史数据进行聚类和归因分析，进而对未来的发展趋势进行预测和判断的。由于未来是由未来的环境所决定的，这种用历史去预测未来的方法本身就具有一定的局限性。而在大数据技术的应用下，人们可以基于多维度数据与某一风险事件之间的相对关系，利用实时的多维度数据对未来进行分析和预测。

2. 改变风险定价的模式

保险公司的承保定价能力是其在同业竞争中的核心竞争力。但一直以来，保险公司对其保险产品所实行的是统一定价原则，很难对客户形成吸引力。在大数据技术的应用下，

保险公司过去的样本精算将升级为全量精算，风险定价模式将发生很大的改变。通过应用大数据技术，传统的保险精算中将引入更多的定价因素，保险公司能够根据客户的特定风险来调整承保定价，不仅能够使客户的差异化需求得到满足，还能使保险公司的承保风险得到降低，从而达到客户和保险公司双方共赢的目的。

1) 增加更多的辅助定价因素

将大数据技术应用于承保定价，能够在其传统的保险产品中增加更多的辅助定价因素，进而帮助保险公司实现对特定客户的个性化风险定价。大数据技术在承保定价中的作用目前在车险和健康险中均有所体现。

在车险领域，基于使用的定价模式已逐渐被保险公司运用在产品创新中。除了获得相关的车型数据、汽车零整比数据、二手车数据以外，保险公司还通过与 4S 店合作获取车辆的保修、保养数据，通过使用车载传感设备收集驾驶员的行驶路线和驾驶习惯数据，进而开发出基于使用的车险计划(UBI，Usage Based Insurance)。在健康险领域，保险公司通过与医院合作掌握客户的健康记录、就诊记录、体测指标、体检报告，甚至是家庭主要成员的医疗记录，通过利用可穿戴设备(如：Jawbone 推出的 Up、Apple 推出的 HealthKit)能够实时监控客户的健康情况(如运动量、睡眠、心跳等)和生活习惯，以弥补生命表对具体的某一客户个体的健康状况和生死概率的判断能力之不足。

2) 根据客户行为的变化进行调整

此外，保险产品的定价调整和客户行为也是相辅相成的，即保险产品的定价是根据客户行为的变化进行调整的。退货运费险的定价模式调整就是典型的例子。

华泰保险于 2010 年和电商平台淘宝合作，针对消费者网上购物所面临的退货风险推出了退货运费险。但该退货运费险在推出后所产生的直接赔付率曾一度高达 93%，其基于客户历史退货情况的产品定价系统也被怀疑是错误的。而造成的这一现象的原因在于，消费者在购买退货运费险后其退货行为变得更加随意，只要有丝毫的不满意都会选择退回其所购物品。因此，华泰保险对其退货运费险的定价系统进行了调整，将包括商品种类和商户的阶段性销售数据等更多的定价因素纳入其定价系统中，进而综合若干数据模型来预测消费者发生退货行为的概率。

从中我们可以看出，基于大数据技术和全局数据的保险产品定价模式可以帮助保险公司在了解客户特点的基础上，设计出满足客户具体保险需求且具有较低风险概率和较高收益的保险产品，进而使保险公司在产品收益、客户体验、风险管理等方面获得优势。

3. 大数据助力保险费率的市场化改革

目前，保险费率形成机制的市场化改革进程在不断加快，意外险、投资连接险、普通型寿险、万能险和非车险等相关领域的费率市场化定价已相继放开，商业车险、分红险的费率市场化定价也即将发令放行，在未来将有更多保险产品的定价权交给高效的市场。保险费率市场化改革的关键在于费率形成机制是科学且有效的。因而在大数据技术的应用背景下，在基于大数法则确定保险产品基准费率的基础上，运用大数据技术为保险产品的附加费率进行定价。

一方面，应由保险监管部门主导构建起公开公正的保险基准费率形成机制，并同时建

立保险基准费率定期测算和发布机制，特别是通过借鉴国际上的成熟经验和模式，设立独立的保险费率厘定机构，进而形成主要保险产品的定价参照基准体系。另一方面，要鼓励保险企业在遵循基准费率的同时，发挥大数据技术在保险产品区域化创新、差异化创新和个性化创新方面的支撑作用，最大限度地处理好保险产品创新与其风险和收益间的关系。

5.2.3 大数据在车险定价中的应用

车险保费的高低一直是车主最关心的话题。在大数据技术应用以前，不论客户驾驶行为的好坏，车险保费的价格基本相当。而随着大数据技术在保险行业的广泛应用，过去优质车主为高风险车主买单的现象将不再出现，基于车主驾驶行为的保费定价模式也将使传统车险的定价模式被完全颠覆。

1. 车险费率厘定的基本模式

通常保险公司在为车险费率进行定价时主要参考两类风险因素：第一类是与机动车辆相关的风险因素，包括品牌、购买价格、使用情况等方面；第二类是与车主相关的风险因素，包括车主的年龄、婚姻状况、职业、驾驶行为等方面。因而我们可以将车险费率的定价模式划分为从车定价模式和从人定价模式。

1) 从车定价模式

在从车定价模式中，保险公司在为投保车辆进行保费厘定时，只考虑与该投保车辆相关的风险因素。这些风险因素包括但不限于：

(1) 投保车辆的种类：可分为客车、货车、摩托车、专用车和拖拉机；

(2) 投保车辆的产地：可分为国产车和进口车；

(3) 投保车辆的使用性质：可分为营业性和非营业性；

(4) 投保车辆的行驶区域：不同区域的车辆在车险费率厘定时也有所不同。

目前，我国车险费率厘定主要采用的就是从车定价模式。该模式具有操作简单的特点，但未考虑与车主相关的风险因素。

2) 从人定价模式

在从人定价模式中，保险公司在为投保车辆进行保费厘定时，主要考虑与该投保车辆的车主相关的风险因素。这些风险因素包括但不限于：

(1) 车主的性别：包括男性和女性；

(2) 车主的年龄层次：可分为青年人、中年人和老年人；

(3) 车主的驾龄：可分为首次领取驾驶证后不足 3 年和首次领取驾驶证后超过 3 年；

(4) 车主的驾驶行为：可分为安全、较安全、一般、较危险和危险。

其中关于车主驾驶行为是依据车主在日常驾驶中的具体驾驶行为数据进行综合判定而来的。从历史统计数据中来看，女性车主的驾驶风险要低于男性车主；中年车主由于具有一定的驾驶经验和较为良好的身体状态，其驾驶风险要低于老年车主和青年车主；首次领取驾驶证不足 3 年的新手车主的驾驶风险高于领取驾驶证超过 3 年的车主；而具有良好驾驶行为和习惯的车主的驾驶风险要低于驾驶行为较差的车主。从人定价模式更加强调车主自身的风险特征，在定价时更加强调个性化。

2. OBD 和 UBI 车险

1) OBD 与车险费率厘定

OBD (On-Board Diagnostics)即车载自动诊断系统，是能够测度和读取机动车辆的运行参数，具有车辆检测、维护、管理等功能的程序系统。OBD 系统能够读取机动车辆发动机、变动箱和 ABS 等的故障码，再通过小型的车载通信设备(GPS 导航仪或者无线通信等)将机动车辆的基本信息、所在位置或者故障码等自动上传到管理平台或设备上。

OBD 系统又被称为 OBD 盒子，在将 OBD 设备插入到机动车辆上的 OBD 插口之后，该设备就能对所检测机动车辆的行驶里程、油耗、发动机转速、故障情况等数据信息进行读取和分析，进而将所分析出的该机动车辆的车辆状况以及驾驶员的行为习惯上传到管理后台上。保险公司可以充分利用 OBD 系统的车辆信息获取、分析和传输功能，了解车主的具体驾驶行为和习惯，对车主的驾驶风险做出精准的判断，从而为车主提供基于其真实驾驶行为的个性化车险费率厘定。

保险公司利用 OBD 系统进行车险费率厘定的数据处理流程如图 5.5 所示。

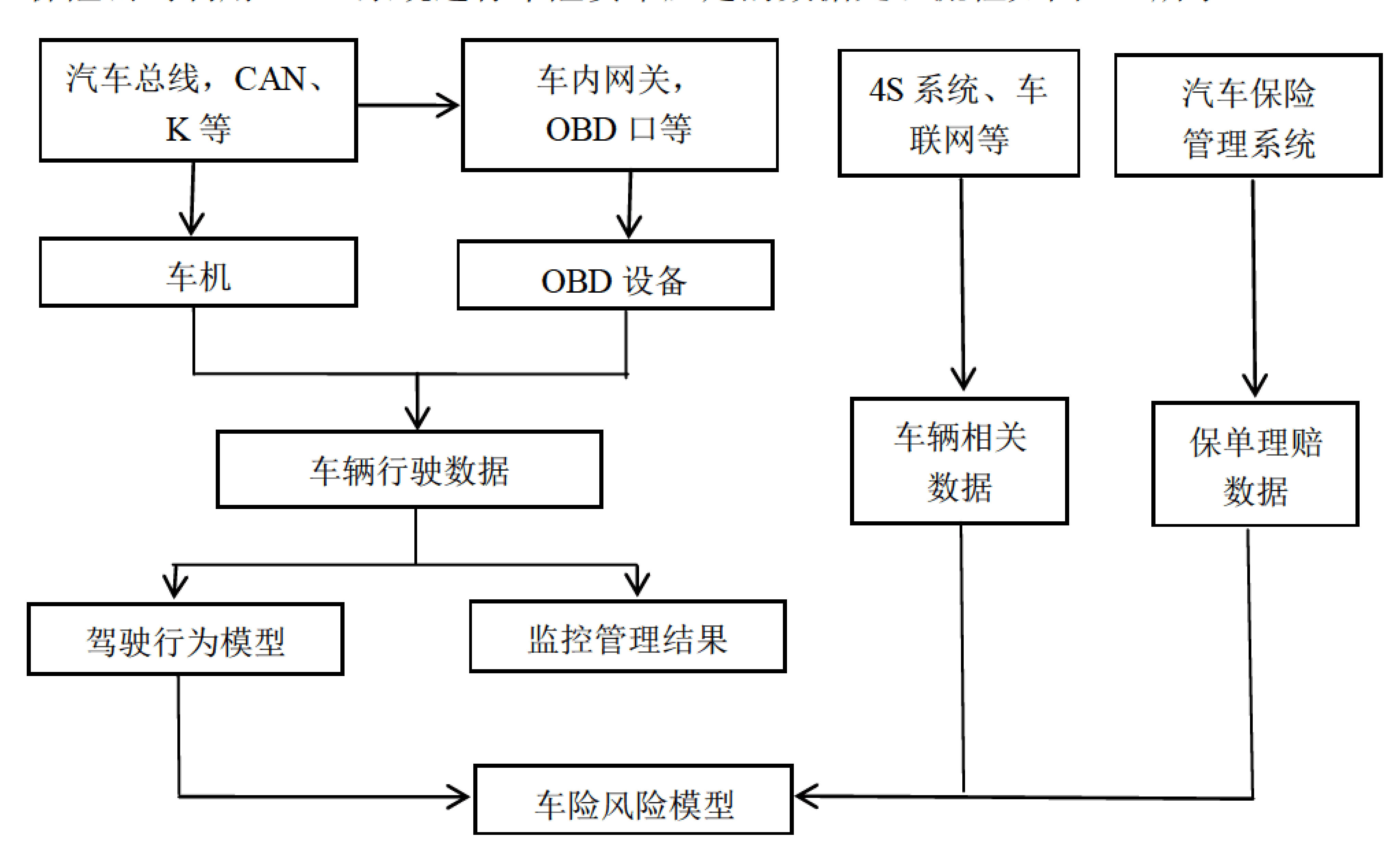


图 5.5 基于 OBD 系统的车险费率厘定

在基于 OBD 系统厘定车险费率的数据处理流程中，车辆行驶数据的采集和分析过程主要由车载终端、采控网关和管理平台 3 个部分组成。其中，①车载终端包括 GPS、CAN 总线的数据采集分析、可视倒车、硬盘 MP5 播放和录像、GPRS 无线数据传输等功能。②采控网关介于硬件终端和上层分析软件之间，具有海量存储、平衡负载、信息交互和预处理等功能。③管理平台一般进行数据的挖掘、清洗以及报表的生成，具体包括安全管理和节能管理等，其中安全管理工作包括远程故障、不良驾驶行为以及车辆部件预警等的诊断，而节能管理工作包括驾驶行为分析、车辆油耗分析、车险匹配分析以及单车运行分析等，此外还设计车辆身份信息以及一键呼叫等功能。因此，一般利用 OBD 系统收集的数

据主要包括超速报警、不良驾驶行为记录(急刹车、急减速、急加速等)、未打转向灯转弯记录、疲劳驾驶管理、出险报警、偷油报警等。

2) UBI 与车险费率厘定

UBI(Usage Based Insurance)是指基于机动车辆驾驶人驾驶行为状况进行个性化保费率厘定的车险。在 UBI 车险中，保险公司将根据实时监测并获取的与驾驶人驾驶行为和习惯相关的各项数据，通过分析和挖掘进而对该投保车辆的驾驶人风险程度进行判断，并将该风险判断的结果应用于车险费率的厘定之中——根据驾驶行为安全性的不同，对拥有安全驾驶行为的投保人给予与其风险程度相匹配的保费优惠，而对具有危险驾驶行为的投保人收取更多与其风险程度相匹配的保费。在 UBI 车险定价中，与驾驶人驾驶行为相关的数据是通过安装在机动车辆上的 OBD 设备获取的。

UBI 车险的应用能够给投保人、保险公司及公共安全带来益处。

对投保人的益处有以下几个方面。

(1) UBI 车险能够为投保人提供更加公平合理的车险费率。尤其是对拥有安全驾驶行为的投保人而言，能够有效地降低其保险支出的负担。

(2) UBI 车险能够帮助驾驶人形成良好的驾驶行为习惯。

(3) 基于 OBD 设备的使用驾驶人可以及时地了解车况并主动地控制风险。

对保险公司的益处有以下几个方面。

(1) UBI 车险使保险公司的车险费率厘定更加科学准确，使其客户满意度和市场影响力得到有效的提高，为其提供了更多的增值效益。

(2) 保险公司在 UBI 车险中能够对投保车辆可能发生的风险进行实时的动态监控，准确了解投保车辆的车况，从而在赔付环节做出合理的赔付，理赔效率得到有效提升。

(3) 基于 OBD 设备在投保车辆上的使用，保险公司的赔付成本也能得到有效地降低，进而实现其经营利润的增加。

而就社会公共安全方面来说，UBI 车险产品的推出和使用能够在一定程度上降低交通事故发生的可能性，有利于社会公共安全的维护。

目前，我国保险市场还没有正式推出 UBI 车险产品，但 UBI 车险产品已经在许多国家的保险市场上出现。在美国车险市场上，出现了基于驾驶里程进行车险费率厘定的 Metronome 项目、基于驾驶行为表现(包括总驾驶里程、日驾驶里程、急刹车次数和时速超过 80 英里/小时的次数等数据指标)进行车险费率厘定的 Allstate 项目等 UBI 车险产品项目。在欧洲车险市场上，也逐渐出现了针对高保费群体的车险 UBI 项目和基于良好驾驶行为给予投保人一定车险费率折扣的 UBI 车险产品。

3. 基于 OBD+UBI 的车险费率厘定

基于 OBD+UBI 的车险费率厘定方法就是以驾驶人驾驶行为为基础，根据驾驶人的不同风险程度确定特定投保人保费水平的差别化车险保费厘定方法。

1) “从车+从人”的定价模式

OBD 设备与 UBI 车险相结合，可以基于投保车辆的车辆状况以及驾驶人的驾驶行为习惯对投保人的车险需求进行风险判断，进而为风险不同的投保人厘定不同的车险费率。

具体来讲，在这一模式下保险公司为确定某个具体投保人的驾驶风险程度，利用 OBD 系统所能收集的相关数据包括：能够反映车辆状况的车辆行驶区域、总行驶里程、日行驶里程、发动机状态等相关数据，以及能够反映驾驶人驾驶行为习惯的急刹车次数、急加速次数、急减速次数等相关数据。在获取数据的基础上，OBD 系统能对该投保人车险需求的风险程度进行量化评判，一般评分越高该投保人的风险程度越低；反之，评分越低该投保人的风险程度就越高。

保险公司在对投保人的车险需求进行费率厘定时，除了要考虑利用 OBD 系统所获取的变动数据外，还会考虑到与投保车辆和驾驶人相关的一些不变因素。这些不变因素包括投保车辆的品牌、车型、出产地、购置价格、车龄等因素，以及驾驶人的年龄、驾龄、性别、健康状况等因素。将不变因素与可变因素相结合、既从车又从人的车险定价方式相较于传统的车险费率厘定方式更加科学合理。

2) 车险费率的厘定方法

基于 OBD+UBI 的车险费率主要是由基础费率和附加费率两部分共同构成的。

基础费率的厘定主要是通过采用传统的费率厘定方法来实现的。在该厘定方法下对投保人进行相关风险的判断，所考虑的是与投保车辆和驾驶人相关的不变因素，即通过对投保车辆的品牌、车型、出产地、购置价格、车龄以及驾驶人的年龄、驾龄、性别、健康状况等因素进行交叉分类，进而确定出该投保人的基础费率。

而附加费率的厘定主要是通过利用内含大数据技术的 OBD 系统对驾驶人的驾驶行为和习惯给出的分数，来确定该投保人应缴纳的车险附加费率。一般评分越高所需缴纳的附加费率就越低；反之，评分越低所需缴纳的附加费率就越高。

在借助大量的数据和车险精算模型厘定出投保人的基础费率和附加费率之后，保险公司根据其所赋予车险基础费率和附加费率的不同权重，计算出该投保人应缴纳的车险费率。

3) 车险保费的支付方式

目前我国保险业还没有出现基于 OBD 和 UBI 的车险产品。因此，由于缺少相当数量的驾驶人驾驶行为数据和评分数据，在该模式下的车险保费支付方式应当采用期初预付当期保费、期末根据投保人相关风险状况多退少补的保费支付方式。

在该支付方式的具体实施过程中，保险公司会在投保人的投保首期以一定的费率优惠鼓励投保人在其投保车辆上安装 OBD 设备，并根据传统车险费率厘定方式来确定投保人应缴纳的保费。之后保险公司根据其利用 OBD 系统对整个投保期内投保人驾驶行为风险的判断，通过保险精算模型计算出该投保人的基础费率和附加费率，进而在给定权重的基础上计算出该投保人实际应缴纳的保费。若该实际应缴保费少于投保人期初已缴保费，保险公司将向该投保人退还多收取的保费；反之，则由投保人补足差额部分。而在之后各期的期初，保险公司将根据上期投保人所实际缴纳的保费预收当期保费，并根据相同方法在期末进行多退少补。

5.2.4 大数据在健康险定价中的应用

随着人口老龄化加速现象的出现，我国所面临的健康和养老挑战越来越严峻。虽然我

国已经初步建立了基本养老、基本医疗等社会保障制度，但相关投入仍然有很大的不足，保障水平依然有限。据有关部门预测，我国健康服务业的规模将在 2020 年突破 8 万亿元，健康和养老服务将成为未来新的经济增长点。随着政策的不断推动以及市场上健康服务需求的进一步释放，商业健康保险将会成为我国医疗保障系统中不可或缺的重要组成部分。

我国健康险的发展起步较晚，相对于人体生理健康变化的周期显得较为短暂，因而我国保险业对于疾病发生率、医疗费用支出率等医疗数据的历史积累较为薄弱。

就医疗信息数据的利用来看，国家层面的人口健康数据应用平台尚未建立；省级层面的人口健康信息平台虽然已陆续开始建设但仅限于卫生系统内部使用；保险业内也尚未建立医疗信息数据的共享系统，与保单相关的大量医疗信息只记录在病历和赔付档案里。从中可以看出我国医疗信息数据的利用程度较低。

这一系列数据运用的问题导致我国健康险产品存在设计不科学、定价不精准、获客困难、医疗费用难管理、道德风险和骗保现象时有发生的问题，而大数据技术的应用能够有效地解决上述问题。下面主要对大数据在健康险定价中的应用进行介绍，大数据在健康险精准营销和欺诈识别中的应用将在之后的小节中进行介绍。

1. 医疗大数据

目前我国的商业健康险可分为团体险和个人险两种。其中，在个人险产品中，大部分是储蓄理财型健康险，而真正意义上的健康险——消费理赔型健康险只占很少的一部分。而导致这一现象出现的原因就在于我国的保险公司对相关医疗费用的估算和控制能力十分有限，且缺乏对相关健康险进行精算定价的数据依据，从而使消费理赔型健康险的设计开发较为困难。

例如，对于肿瘤类的大病保险，政府医保基于保基本原则只能支付其治疗费用中的一部分，且报销范围不涵盖当前市场上治疗效果显著但价格昂贵的靶向性生物试剂，使肿瘤患者及其家庭面临着较大的治疗负担。对于这一保险市场中的空白，虽然有许多保险看好这一市场机会，但碍于其无法准确掌握治疗肿瘤疾病的实际医疗费用进而对该保险产品进行合理定价，只能对该市场机会望而却步。

1) 我国的医疗信息化建设

2014 年国家卫计委提出中国卫生信息化建设的顶层设计规划——“4631—2 工程”。

该工程由以下几个部分组成。

(1) “4”代表 4 级卫生信息平台，分别为国家级人口健康管理平台、省级人口健康信息平台、地市级人口健康区域信息平台以及区县级人口健康区域信息平台。

(2) “6”代表 6 项业务应用，分别为公共卫生、医疗服务、医疗保障、药品管理、计划生育和综合管理。

(3) “3”代表 3 个基础数据库，分别为电子健康档案数据库、电子病历数据库和全员人口个案数据库。

(4) “1”代表 1 个融合网络，即人口健康统一网络。

(5) “2”代表 2 个信息体系，即人口健康信息标准体系和信息安全防护体系。

(6) 相关医疗信息的来源包括：基于电子病历的医院信息系统、基层医疗卫生管理信息系统、医疗健康公共服务系统和计划生育信息系统。

从中我们可以看出，“4631—2 工程”是致力于打造全方位、立体化的国家医疗卫生信息资源体系的国家顶层设计规划。该工程一经建成并投入使用将会使我国医疗大数据巨大的使用价值得到充分的发挥和体现，如图 5.6 所示。

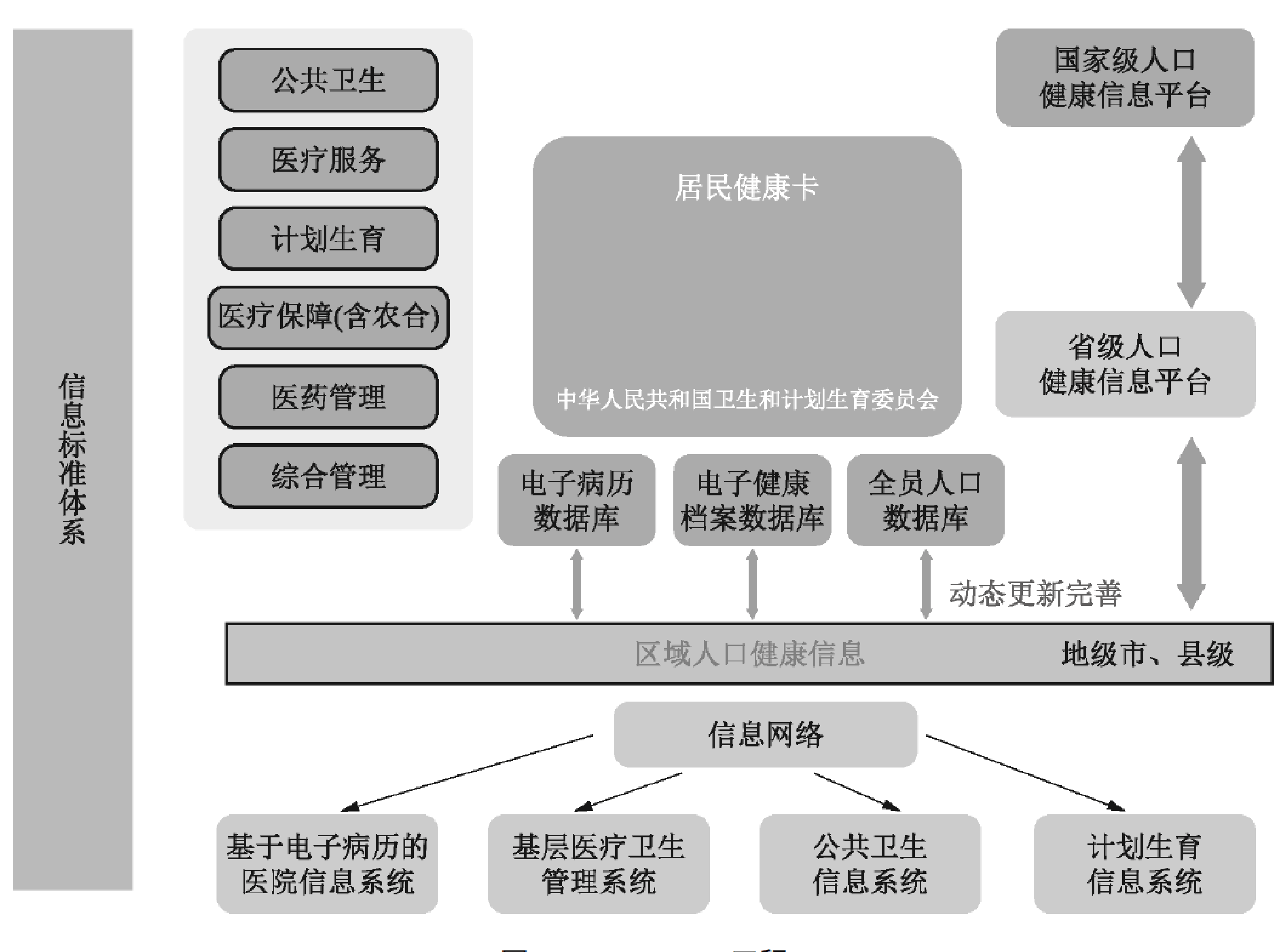


图 5.6　4631—2 工程

2)　医疗数据库体系

在卫计委“4631—2 工程”的 3 个基础数据库——电子健康档案数据库、电子病历数据库和全员人口个案数据库中，电子健康档案数据库和电子病历数据库是最为基础和重要的医疗信息数据库，如图 5.7 所示。

(1)　医院是医疗大数据获取的关键来源。

据有关数据显示，近 80%的药品销售是在医院中实现的，而患者接受医疗服务的过程是在医院中进行的，医保消费的支出也主要配置于医院之中。从而可以看出，与个人相关的医药、医疗和医保数据都主要汇集于医院之中，医院是医疗大数据获取的关键来源。

电子病历是医院信息平台的核心，因而电子病历系统是医院信息系统中最为重要的医疗数据系统。

(2)　居民健康档案是个人健康数据存储的主要渠道。

居民健康档案中所存储的相关数据如图 5.8 所示。

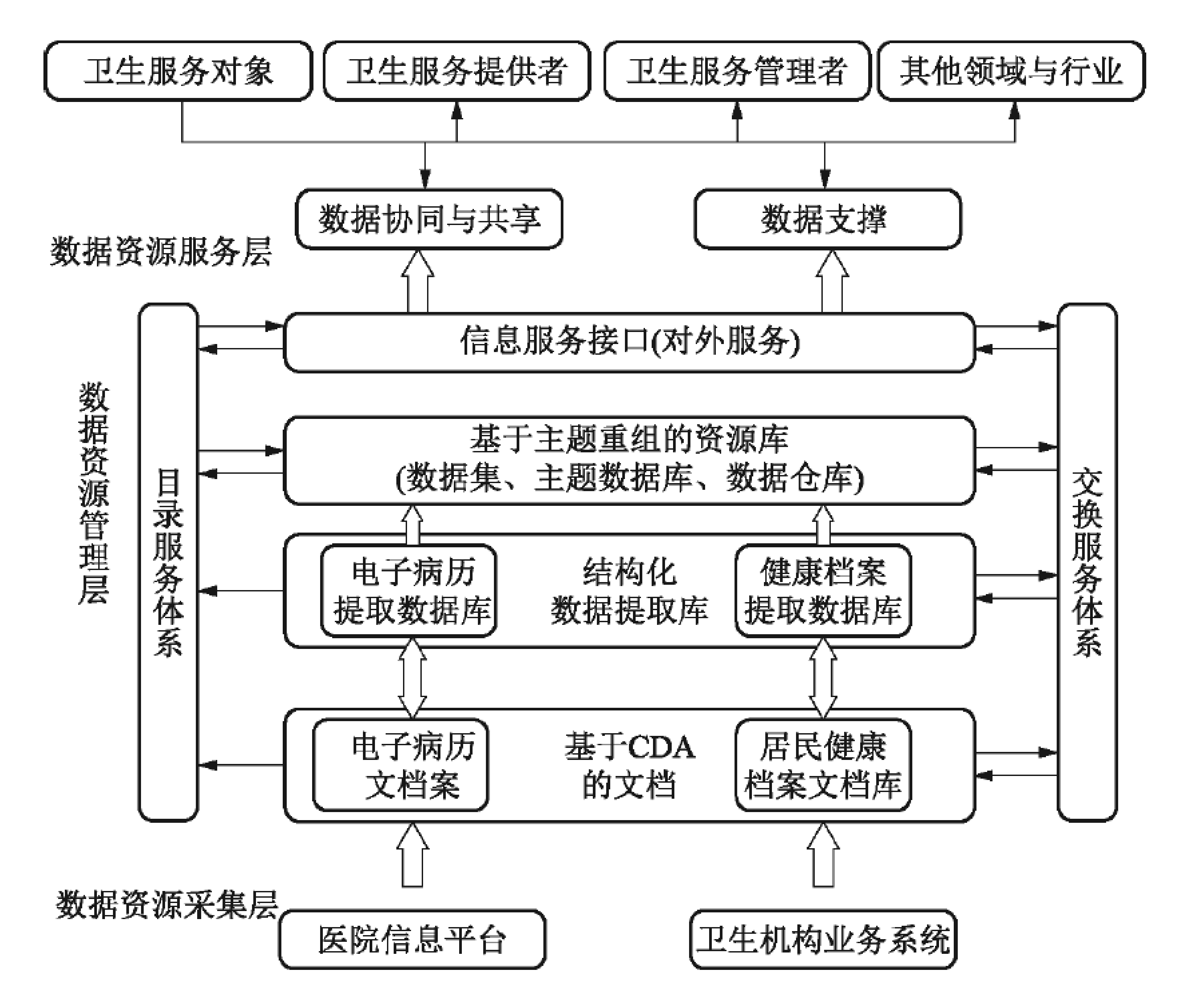


图 5.7 医疗数据库体系

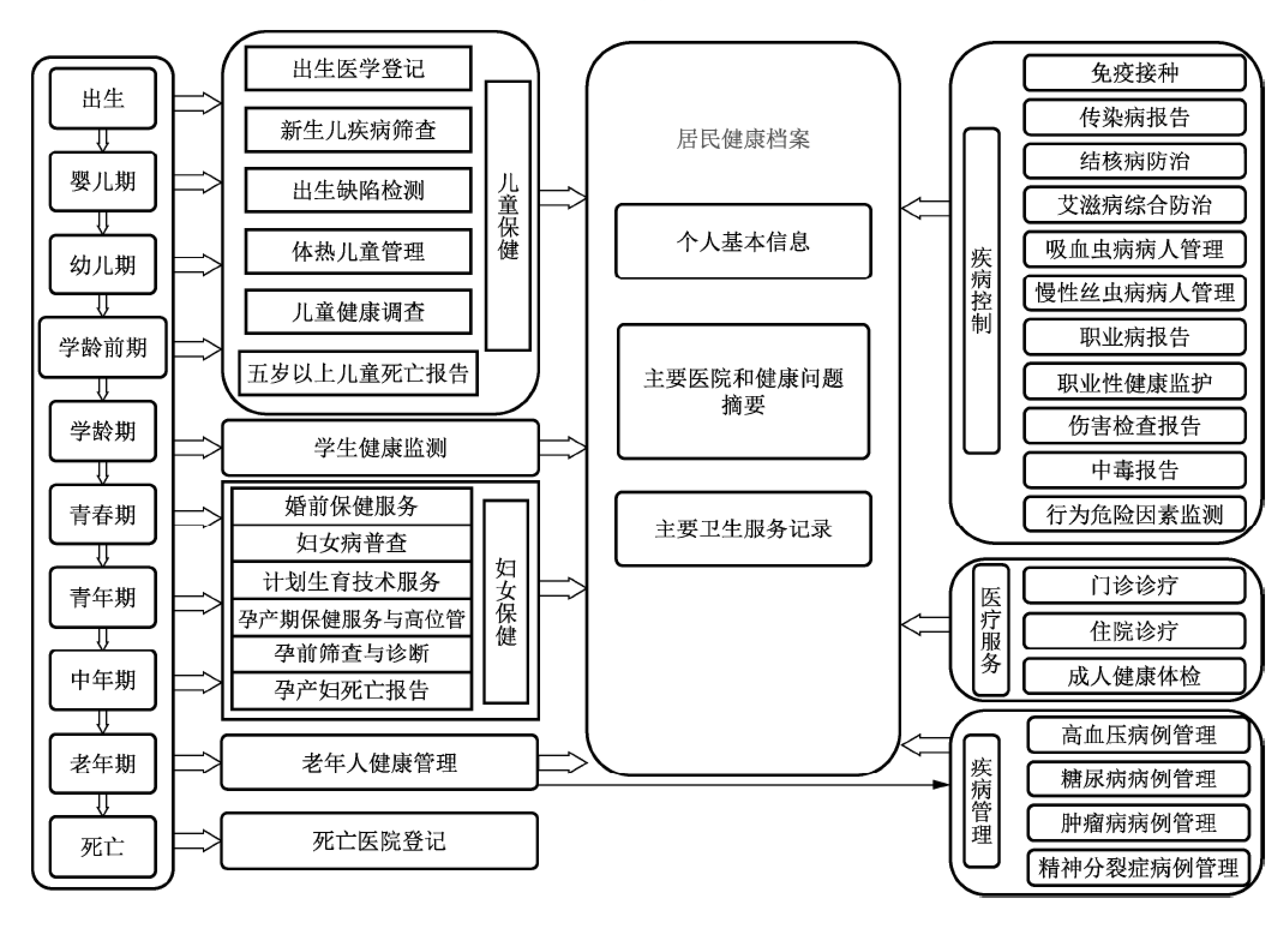


图 5.8 居民健康档案中存储的数据

居民健康档案是区域医疗信息平台的核心，其中所存储的数据不仅包括与居民个人相关的电子病历数据，同时还包括个人的公共卫生信息。个人的基本信息、主要就诊医院和相关健康问题摘要以及其所接受的主要卫生服务记录都被记录在居民健康档案之中。借助居民健康档案的建立和运作，与每个居民个体相关的医疗健康数据都被有效地存储，并在合理合法的范围内交换和流通。

3) 可穿戴设备的应用

随着科学技术的快速进步和发展，获取医疗大数据的途径并不仅仅局限于政府所开发的医疗数据库体系，可穿戴设备的应用为保险公司获取其被保险人健康信息提供了另一种有效途径。在大数据技术的应用下，保险公司可以通过分析可穿戴设备所收集的被保险人各项健康指标数据实时了解被保险人的身体状况，并从众多被保险人的健康数据中分析出健体与弱体之间的差异。

目前可穿戴设备的应用设计已日趋成熟，各类可穿戴设备相继出现在市场当中。例如，蓝牙耳机和扬声器厂商 Jawbone 推出了其可穿戴设备——UP，国内的小米公司推出了其可穿戴设备——小米手环，苹果公司在其 iOS8 系统中推出了名为 HealthKit 的集成应用。

南非最大的健康险公司——Discovery 公司推出了 Vitality(健行天下)健康促进计划，该计划致力于通过建立科学的健康管理和激励体系，鼓励其被保险人关注自身健康，并以恰当方式对其被保险人的健康行为和饮食习惯进行干预。Discovery 公司以该健康促进计划为基础建立了保费的活力优化系统，被保险人自身的活力状态会对他们的实际保费产生影响，且被保险人的活力状态越好，其所能享受到的增值服务奖励相应地也就越丰富。在对被保险人的活力状态进行测算时，Discovery 公司采用了线上与线下渠道相结合的方式获取被保险人的相关健康数据，其中线上渠道是指通过利用 Withings 推出的可穿戴设备获取被保险人的活动数据和健康数据，而线下渠道是与健身机构进行合作。

2. 大数据与健康险的产品设计

健康险产品设计必须兼顾社会伦理和保险成本，通常包括确定所提供的服务和进行产品定价两个方面的工作。

1) 健康险提供的服务

在考虑健康险产品所提供的服务时，会对以下内容进行确定。

(1) 单个被保险人在本期间内发生的医疗费用支出，本产品能够负担多大的比例。

(2) 在不同的健康状态下，被保险人未来罹患某种疾病的概率，以及各类疾病的平均诊治费用。

(3) 单病种的报销额度，即被保险人罹患某种疾病时本产品能为其报销多少手术费用、医药费用以及住院费用。

在对第(2)项内容进行确定时，保险公司可以利用大数据技术从海量电子病历数据中，计算出各类疾病的平均诊治费用，并且通过跟踪多位患者的病情发展状况，计算出疾病转化的概率。

2) 健康险产品定价

保险产品定价的主要依据是理赔标的发生概率。在大数据应用以前，测算理赔标的发

生概率所利用的大部分数据都是来源于行业内的历史数据和统计数据。随着科学技术的不断进步，现今无论是疾病的诊断方法还是治疗模式都发生了巨大变化，历史数据已不再具有代表性，而且传统的数据统计方式已经过时。例如，在重大疾病险中，心肌梗死的冠状动脉造影早已是诊断该疾病的最佳标准，但在相应健康险产品定价中仍沿用老的诊断标准，造成其定价失准。

而在大数据技术的应用下，健康险产品的定价将更为精准。大数据思维认为，小样本数据会使误差加大，依靠误差较大的数据无法设计出接近真实概率的产品。因而要通过利用大数据技术对海量数据进行分析挖掘，保证其产品定价与客户投保的需求偏好相一致，避免由于定价过高而无法被潜在投保人认可和接受，或由于定价过低使保费难以覆盖风险，进而使保险公司自身产生亏损。

【案例 5.1】 UBI 车险产品和 UBI 车险服务

1. UBI 车险产品

英国 Aviva 保险公司针对年轻司机需要负担高于其风险水平的保费这一现象，借助大数据分析，开发了基于驾驶人驾驶行为的驾驶风险预测模型，从而实现了个性化定价。这一举措不仅改善了投保驾驶人驾驶习惯，同时也为公司削减了一定的运营成本。Aviva 保险公司不仅对客户个人信息、车辆信息和使用情况、驾驶历史等数据进行收集，还引入车载设备，以通过手机 APP 来监控驾驶人在起步后行驶 200 英里的驾驶状态。Aviva 保险公司根据驾驶人驾驶行为(如加速、刹车和拐弯时的频率和程度)的数据记录，从中分析出该驾驶人的驾驶风险并对其进行定价——确定个性化的保费，并向该驾驶人提供相应的保险服务。同时 Aviva 保险公司还为安全驾驶者提供最高达 20%的保费折扣。实施后的相关数据显示其被保驾驶人的驾驶安全状况有所改善，Aviva 保险公司这一新商业模式也为其赢得了更高的客户满意度和有所降低的客户流失率。

Metromile 保险公司借助汽车监控设备的使用对其车险定价模式进行了调整，从而实现了按驾驶里程收费。它的里程定价模式是基于车载汽车监控设备的技术，通过客户安装的设备追踪投保车辆的行驶里程进而为其确定应缴纳的保费。Metromile 保险公司的投保人只需每月支付 15～40 美元的固定费用以及 2～6 美分/英里的使用费即可。这一款车险产品并不考量驾驶人如何开车，而仅关心投保车辆所行驶距离。Metromile 保险公司的这款保险产品在行驶里程不多且尚未被充分服务的车险细分市场中有很大空间。平均计算来看，这款保险产品可为年行驶里程在 10000 英里的驾驶者节省 40%的保费。

2. UBI 车险服务

美国 Liberty Mutual 保险公司为企业客户的大型车队提供 GPS 跟踪监控设备。企业客户将该设备安装在其所有的汽车上，可通过该 GPS 跟踪监控设备回传的里程数、行车时速、加速和刹车情况以及车辆所处位置等与投保车辆相关的数据信息，进而帮助投保人对其车队进行监控并帮助车队司机改善其驾驶习惯，并在此基础上进一步开展车辆安全管理，从而有效地对相关风险进行控制。

英国 Insurethebox 保险公司将含有 GPS、运动传感器、SIM 卡和电脑软件的车载盒子装在汽车上，通过 GPS 技术追踪定位失窃车辆，协助客户找回。当该车载盒子检测到车辆

发生撞击或意外事故时，Insurethebox 保险公司会及时与客户通话，对客户人身安全进行核实；在特殊的紧急情况下，Insurethebox 保险公司还会呼叫应急救援部门参与事故救援。而车载盒子里所存储的数据也可用于协助公司对投保车辆的毁损情况进行分析。

(资料来源：大数据及车联网在车险中的应用和案例)

【案例 5.2】 大数据下的健康险

众安在线人寿保险公司推出了其大数据智能健康险产品——步步保，这是众安在线人寿保险公司和小米运动、乐动力 APP 合作推出的保险产品。客户(即被保险人)投保时，系统会根据其历史运动情况以及预期运动目标，向其推荐不同保额档位的重大疾病保险保障(目前分 20 万元、15 万元、10 万元三档)，用户历史平均步数越多，推荐保额就越高，最高可换取 20 万元重疾保障；其中，如果被保险人在参加健康计划前 30 天的平均步数达到 5000 步，则被推荐 10 万元保额重大疾病保险保障。在申请加入健康计划后，申请日的次日会作为每月的固定结算日，只要每天运动步数达到 10 000 步，下月结算时其保费就可以多免费 1 天。即保单生效后，用户每天运动的步数越多，下个月需要缴纳的保费就越少。对于这款以运动数据作为其实际承保定价依据的保险产品，众安在线人寿保险公司称其为“国内首款与可穿戴设备及运动大数据结合的健康管理计划”，并表示未来将会接入更多可穿戴设备和运动 APP，进而通过覆盖更多的运动人群以实现其产品定价和规模优势的双提升。

(资料来源：凤凰财经)

5.3 精准营销

大数据能够帮助保险公司收集海量且多样的客户数据，使保险公司能够基于大数据的分析结果找出不同客户的潜在保险需求，进而将不同的保险产品恰当地推荐给有该产品潜在需求的特定客户。因而在大数据技术应用的背景下，保险公司的营销不再是以同一个广告内容和营销手段对所有的潜在客户群体进行营销，而是针对具有不同保险需求特征的细分客户群体进行有针对性的营销。随着移动互联网技术的快速发展和智能移动设备的日益普及，各类手机应用客户端所收集的客户各类操作行为、人们在其社交媒体上分享的文字、图片、视频等都可以成为了解和刻画客户的重要数据，借助大数据技术对这些数据进行采集和分析，保险公司可以准确地了解客户的特点和需要，为数据价值的商业运用提供基础。

5.3.1 保险精准营销

1. 保险精准营销的概念和步骤

大数据背景下的保险精准营销，是指保险公司在可量化的数据基础上对单一客户的消费模式和特点进行分析和归纳，对其客户群体进行划分，进而精准地找到其目标客户并精准地向目标客户开展营销活动，以提高其营销效率的过程。

保险公司进行精准营销的步骤如下。

1)　客户信息采集

了解客户的基本信息和行为偏好是精准营销实现的基础。因而保险公司不仅要对其内部掌握的客户基本信息(如客户的年龄、性别、家庭成员状况、学历、职业、收入、资产持有状况等)加以利用，还要加强与网络购物平台、网络社交平台以及其他掌握客户数据的第三方进行合作，从而获取更多的客户行为信息。

2)　用户数据分析

在前一步骤对海量客户数据进行收集的基础上，保险公司要利用大数据理论和分析模型对所收集的客户数据进行相应的分析和挖掘，从而实现对客户特征和客户行为的精准刻画和描述。

3)　结果分析解读

基于上一步骤的分析结果，保险公司可以对其所面临的众多客户进行有效的细分，并对每一细分客户群体的保险需求进行分析和判断，进而为其匹配恰当的保险产品。

4)　实施营销

在前述步骤有效实施的基础之上，保险公司要根据其每一细分客户群体的偏好特征以及相应保险产品的主要特点，制定出最佳的营销方案，进而在最佳的时间向特定细分客户群体进行营销。

2. 传统保险营销的不足

1)　市场细分不够充分

虽然我国的保险公司对市场进行了一定的细分但不够充分，许多保险公司在进行市场细分时都忽视了保险市场需求在地域之间的差异、群体层次之间的差异、城乡之间的差异以及收入水平的差异。这也导致传统保险营销实质上是广撒网式营销，营销效率较低。

2)　对客户需求不够重视

由于传统保险营销模式缺乏针对性，且保险公司所推出的保险产品又具有同质化的特征，导致同一客户就同一风险标的在同一家保险公司中重复投保的现象出现，这无疑会使客户丧失对保险公司的信心。而且客户多样化的潜在保险需求也无法得到有效的满足。

3)　适应市场的速度较慢

在传统的保险营销过程中，保险公司对潜在市场需求的变化不够敏感；即便是发现了潜在市场需求，碍于其烦琐复杂、效率低下的产品设计过程，其最终所推出的保险产品已无法与客户新的潜在保险需求相匹配。

4)　重短期利润、轻长期服务

在传统的保险营销模式下，保险代理人为了追求更高的佣金收入会向客户推销价格上更具吸引力而与客户实际保险需求不相匹配的保险产品，使有保险意愿的优质客户的保险需求无法得到应有的满足，进而使客户对保险公司的忠诚度和依赖度减少，造成客户流失。

3. 大数据与保险营销环节的契合

我们可以将保险营销的过程分为客户接触、客户联系、客户赢取 3 个环节。在传统的保险营销模式中，保险公司的营销渠道主要包括保险公司自有的销售团队以及各种形式的

保险中介。其中，保险中介作为保险公司与客户之间的媒介，能够利用自身的规模优势降低协调成本，进而为保险公司带来经济价值。在大数据技术应用的背景下，保险营销将通过构建数据产生、数据采集和传输、数据处理应用 3 个环节来替代传统的保险营销过程。其中，数据产生环节与传统营销过程中的客户接触环节相对应，数据采集和传输环节与传统营销过程中的客户联系环节相对应，数据处理应用环节则与传统营销过程中的客户赢取环节相对应(见图 5.9)。

1) 数据产生环节与客户接触环节的契合

在传统的保险营销中，客户接触环节主要是通过保险业务人员来完成的。而在大数据时代，保险公司可以通过数据接触客户，进而实现对客户及其行为习惯实时的分析和预测。即在接触客户之时就有相关客户数据产生。而各类智能设备的推出和普及，使保险公司能够实时地与客户进行交流和沟通，并实时地掌握客户的各项特征。

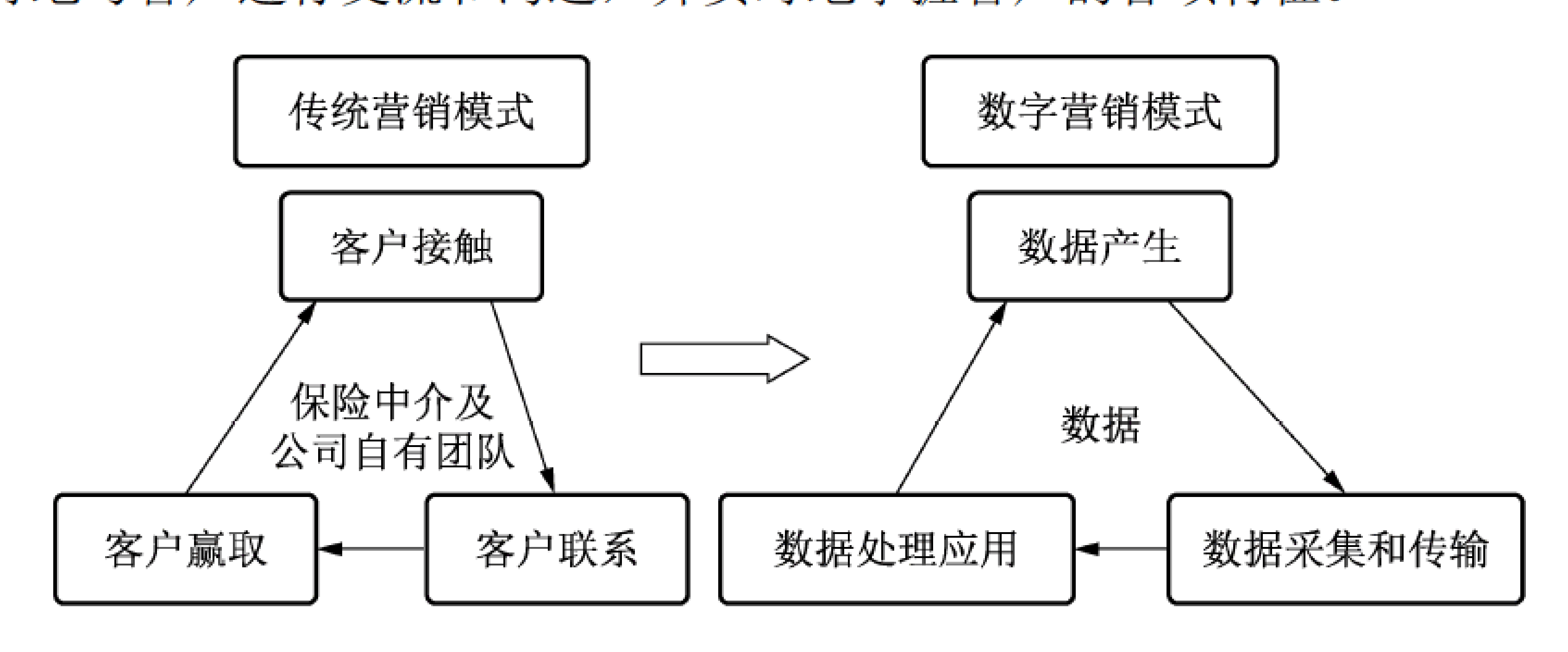


图 5.9 大数据与保险营销环节的契合

2) 数据采集和传输环节与客户联系环节的契合

保险公司在与客户进行接触后需要跟客户进行有效的沟通互动，即与客户之间产生联系。随着近年来通信技术的快速发展，保险公司对客户数据的采集和传输效率得到大幅提高。保险公司与客户的沟通也更加的简单和快捷，二者之间互相传达的信息也更加的直观和全面。

在这一过程中，不仅客户能够享受到优质高效的保险服务，进而与保险公司保持更加紧密的联系；而且保险公司也可以通过高效的客户联系在短时间内聚集大量的客户。

3) 数据处理应用环节与客户赢取环节的契合

客户的相关需求被传递到保险公司后，保险公司需要通过快速且优质的回应来赢得客户。而在回应客户的过程中离不开数据的处理和应用，因为只有高效地对全方位的海量数据进行处理和应用，才能有效地回应客户。大数据技术和云计算技术的结合使大量客户数据在到达保险公司后可以被有效地利用，进而使针对具体客户的评估、预测以及动态定价得以实现，为保险公司对客户保险需求的高效回应提供了保证。数据的处理应用环节使保险公司形成了有效的客户互动循环，进而帮助其赢取更多的客户。

5.3.2 大数据与保险精准营销

精准营销的实现基础是精确定位。精——即科学的细分，市场、客户和服务的细分都

要精；准——即准确的把握，包括信息的采集和分析、公司品牌的定位以及产品服务的投放都要准。保险公司通过利用大数据技术能够实现对市场的有效细分，对其保险产品的目标客户进行识别，在对目标客户的分布特征、信息来源和购买倾向进行分析的基础上进行针对性营销。例如，保险公司可以通过运用新型营销方式为日均手机上网时间超过 6 小时的客户推荐健康保障险、手机意外险，为有吸烟嗜好、经常应酬、爱吃肉食的客户推荐防癌险、重大疾病保险，为使用高端智能手机的客户推荐碎屏险，为在旅行服务平台上消费的客户推荐旅游相关保险产品等。

1. 大数据下的新保险营销方式

保险公司在大数据背景下对客户保险需求进行分析时，相关的数据来源不再仅限于保险行业内部的保险客户数据，还包括保险体系以外的数据，包括与保险公司进行合作的第三方自身所积累的客户数据、网络社交平台所积累的客户数据等。这些外部数据对保险公司的市场细分具有极大的利用价值。在对客户信息和客户行为的相关数据进行深度分析的基础上，保险公司能有效地预测客户需求、挖掘潜在客户并向其推荐适当的保险产品，从而实现个性化的精准营销。大数据下的新保险营销方式有以下几种。

1) 搜索引擎营销

搜索引擎营销(SEM)，是指在搜索引擎平台上所进行的网络营销。保险公司通过与搜索引擎平台合作，利用大数据技术了解该搜索引擎用户个体之间不同的搜索行为特征以及其主要关注焦点，进而在用户检索信息的间隙有针对性地将本公司适宜该用户的保险产品及服务信息传递给该目标用户，进而保险公司可以及时获取到在相关检索结果页面查看其所展示保险产品和服务的用户信息，并及时地向这些潜在客户实施营销策略。与传统保险营销方式相比，搜索引擎营销目前已成为保险公司获取客户、进行精准营销最为直接有效的方式之一，其优势主要表现为以下几个方面。

(1) 突破时空限制，营销对象广泛。

(2) 广告投放精准，且具有针对性。

(3) 信息传播速度快，营销效率较高。

(4) 可以对营销推广的效果进行监控。

2) 微信保险营销

随着移动互联网技术的快速发展，有着强大用户基础的微信平台也在不断开发出更多的服务功能。正基于此，许多保险公司都开始在微信平台上提供保险服务。通过利用大数据技术，保险公司可以有效地获取和了解其在微信平台中的潜在客户，从而向潜在客户进行精准营销。潜在客户可以通过保险公司在微信平台上的服务窗口及时接收到与其保险需求相符的保险产品营销信息，对其感兴趣的保险产品可以直接进行投保并完成相应的支付；存量客户可以通过保险公司在微信平台上的服务窗口接受到便捷的保险服务，如保单信息查询、保单变更、网点查找、客户投诉、理赔咨询、理赔进度查询、快速赔款、报案注销等。保险公司通过微信平台进行精准营销可以使客户的黏性和活跃度得到有效的提高。

3) 微博保险营销

微博作为当前另一大主流网络社交平台，在营销方面也具有极高的利用价值，因而许多保险公司都开通了其微博服务账号。微博在信息传递方面具有很大优势。保险公司通过

其微博服务账号能够及时、直观地向公众发布其保险产品和服务的营销信息，开展多种多样的网络营销活动，并能实时地与其客户进行良好的互动。此外，具有社会化媒体特点的微博会实时发布相关热点事件，保险公司可以借助热点事件巧妙且及时地进行保险营销，从而吸引更多的用户关注，与更多的用户达到共鸣，使其与潜在客户之间的距离不断被拉近，从而达到高效的营销效果。

4) P2P 保险营销

P2P(Peer-to-Peer)保险营销模式，是指在人与人之间的社交关系基础之上所产生的互助保险模式。这不仅是一种新的保险营销模式，更是一种新型的保险存在方式。就发展历史来看，保险最初就是从小团体成员之间所进行的风险管理互助开始的，后来逐渐从基于单纯的人际关系演化为基于合同关系的风险管理互助，保险公司也由此产生。在大数据的时代背景下，最初的互助保险形式又将会回归到风险管理领域，但与最初的互助保险所不同的是，互联网以及大数据的经营理念和相关技术被融入其中，即产生基于互联网的“众保”保险模式。国内外在 P2P 保险领域已经有所实践，如德国的 Friendsurance 公司，我国的“抗癌公社”和泛华保险服务集团推出的“e 互助”等。

2. 大数据下精准营销的实现路径

1) 营销理念的变革

在传统的保险营销理念之下，保险公司已习惯于借助有限且粗劣的营销数据去进行相应产品和服务的营销，对营销成本的控制和营销效果的评价标准都相对较弱。在大数据应用的背景下，传统保险营销的低效率将不复存在，保险公司应与时俱进地积极获取与营销相关的数据信息，了解和洞察其潜在客户，进而以精细且准确的营销方式对客户进行营销，从而使其客户转化率有效提高。

2) 对差异化需求进行分析

(1) 建立数据库。

客户数据的采集和分析是精准营销的基础，因而保险公司应当对多种信息获取渠道进行灵活运用，将其所掌握的行业内部数据和分散于其他领域的外部数据进行汇集，在利用大数据技术对这些数据进行分类和转化后，将这些数据存储于其数据库之中。拥有独立且成熟的数据库是保险公司进行精准营销的前提。

(2) 分析差异化需求。

大数据背景下的保险精准营销更加关注客户差异化的潜在保险需求。保险公司借助大数据技术对其所掌握的与客户相关的海量数据进行分析和挖掘，能够对每一客户个体的潜在保险需求进行预测，从而在合适的时间将合适的保险产品和服务推荐给合适的客户。对客户差异化的保险需求进行预测是保险公司精准营销的关键。

3) 营销模式的全面精准化

(1) 开发精准化的产品。

保险公司基于其所获取的海量数据，以客户的保险需求为出发点设计相应的保险产品和服务，即为其所提供的保险产品和服务增添差异化的市场价值，以迎合相应客户的保险需求。在对客户保险需求进行洞察的基础上推出精准化的保险产品和服务是经济且高

效的。

(2) 制定精准化的价格。

承保定价决定着保险公司的盈利水平，因而保险公司应当利用大数据技术对其所提供的精准化保险产品和服务进行精准化的定价，向不同的投保客户收取与其风险水平相匹配的保费，进而在保险市场中的供给与需求之间找到利润的平衡点。

(3) 进行精准化的营销沟通。

保险公司通过分析和挖掘其所掌握的海量客户数据可以了解客户的兴趣爱好和行为习惯，进而以投其所好的营销内容和营销方式在恰当的时间与潜在目标客户进行营销互动，即对潜在目标客户进行精准化的营销沟通。

(4) 建立增值的服务体系。

保险公司可以依托其所掌握的海量客户数据向客户提供更多与其保险产品和服务相关的增值服务，进而在对客户的特定风险进行管理的同时，以人性化的服务提高其客户的黏性，赢得更多市场竞争优势。

4) 精准营销的效果反馈

保险公司还应当借助大数据技术对其精准营销活动的效果进行评估和反馈，以帮助其改进现有精准营销活动中的不足，从而使其下一阶段的保险营销活动更为精准。

5.3.3 组建垂直平台生态圈

伴随着移动互联网和大数据等新技术的快速发展，保险公司作为保险价值链中的关键一环正在积极地引入数字化技术以优化客户在其所提供保险服务中体验，提升其自身的精准营销能力、风险管理能力和客户影响力。其中一些具有一定实力的保险公司已开始尝试通过建立数字平台来整合保险价值链上的所有相关服务，从而构建起平台生态圈。

1. 平台生态圈的概念

平台生态圈是指商业活动的各利益相关者通过共同建立一个数字化价值平台，将价值链上各利益相关者所掌握的数据、所提供的产品和服务联系起来，以客户一系列的核心需求为出发点向客户提供组合服务，进而达到优于利益各方单独提供产品和服务的经济效果。其中，数字化的价值平台可以是操作系统、应用程序商店、社交网站或其他形式。

平台生态圈与传统业务模式的不同体现在以下 3 个方面。

1) 以客户需求为中心

平台生态圈以客户需求为中心而非以产品和服务为中心，相关利益各方将根据客户需求的不同提供具有差异化的产品和服务，使客户的相关需求均能在该生态圈中被满足。

2) 以数据平台为支柱

与客户相关的信息数据被存储在数字化价值平台之中，并且为平台生态圈的各方利益相关者所共享，以帮助他们准确地捕获客户需求并对其所提供的产品和服务进行优化，进而提升客户体验。

3) 由多方利益相关者组成

平台生态圈由多个身处不同行业的企业组成，客户的相关需求将通过各方参与者之间的业务竞合得以满足。

对平台生态圈的构建方来说，构建平台生态圈的根本目的在于吸引更多的潜在客户、挖掘存量客户的新需求，在扩大企业业务规模或业务范围的同时，实现企业品牌与盈利能力的双提升。

2. 构建垂直平台生态圈的动因

保险公司构建垂直平台生态圈的动因有以下几个方面。

1) 获取更多客户数据

通过构建平台生态圈，保险公司可以从其他参与者处获取到更多行业外部的客户数据；并能在与客户高频率的沟通互动中，提升客户的忠诚度。

2) 实现精准营销

保险公司通过对其在平台生态圈中所获取的客户数据进行有效的数据挖掘，能够实现对客户的细致筛选，并能根据筛选结果有针对性地将客户迁移到其他的产品和服务中去，进而实现精准的客户迁移和市场营销，使其客户贡献度得以提升。

3) 提供更多增值服务

保险公司通过利用其在平台生态圈中所获取的海量客户数据，不仅能够对其现有产品和服务进行优化，还能为其客户提供更多具有针对性的增值服务。

3. 平台生态圈的构建

平台生态圈的构建是复杂的，因此保险公司在打造其保险生态系统时需要完成的工作包括但不限于以下几个方面。

1) 充分了解自身的地位、优势和劣势

保险公司需要对其在价值链与市场中的地位、优势和劣势有充分的认识，进而明确其与其他利益相关者之间的合作模式。即保险公司通过回答本公司能够提供怎样的保险产品和服务、本公司拥有哪些数据、哪些公司需要本公司的数据、本公司需要哪些外部数据来支持产品和服务的优化等一系列问题，能够对其在生态圈中的合作内容和合作方式进行确定。

2) 有效选择合作方

在明确合作内容和合作方式的基础上，保险公司要与价值链上其他行业的具体合作方进行选择。其中，合作方既可以是其他公司，也可以是同一集团内的其他子公司。保险公司通过与潜在合作对象进行接触和沟通，从而确定出其最终具体的合作方。

3) 对平台生态圈进行快速试错

在与相关合作方构建起平台生态圈后，保险公司应从中选择某一个或几个产品和服务作为测试对象，对所构建的平台生态圈进行快速试错，在与合作方相互磨合的过程中完成对测试对象的合理评估，进而对其平台生态圈进行进一步的优化。

【案例 5.3】 保险公司与平台生态圈

德国安联保险公司为实现规模效益和技能互补，与德意志电信展开合作。安联保险公司基于对双方技术优势的利用，通过构建生态系统为其零售客户与企业客户提供独特的产品和服务。

针对其零售客户，安联保险公司与德意志电信协作开发了数字化的“联网之家”服务，该服务是高科技技术与援助服务和保险服务有机结合的个性化增值服务，客户可以利用传感器和智能手机实现对自己家的实时监控。一旦家中发生意外如水管爆裂，传感器不仅会自动通过客户的智能手机通知客户，还会第一时间通知安联保险公司的紧急援助部门。

而对企业客户，安联保险公司与德意志电信合作推出具有全面性的网络安全解决方案以及与之相匹配的保险产品，继而向其企业客户提供个性化的网络服务和保险服务的产品组合。例如，德意志电信的先进网络防御系统与安联网络的防护保险产品相结合，在为客户提供智能化网络防御系统的基础上，还为其提供了最高承保额达 5000 万欧元的保单。

与此同时，安联还积极开展安联全球合作伙伴项目(AWP)，以实现其生态系统的构建。安联全球合作伙伴项目包括安联全球救援和全球汽车、安联全球护理以及安联法国国际健康 3 项内容，能够帮助安联保险公司实现其与交通、医疗与健康等领域之间的融通与协调，进而为其客户提供更为卓越的增值服务。例如，该项目中的安联全球救援和全球汽车子项目：一方面，安联保险公司通过与汽车厂商签订合同，进而为多个品牌的购车客户提供相应的车辆保险产品服务。另一方面，安联保险利用其庞大且有效的服务供应商网络，为其客户提供汽车道路救援服务。在提供道路救援服务的过程中，安联保险公司还通过运用大数据技术对事故的发生进行充分的挖掘和分析，并将分析结果应用于优化其保险产品和服务，进而为其创造更大效益。

(资料来源：《互联网+时代大数据改良与改革中国保险业》之五技术引发商业模式新变革)

5.3.4 大数据精准营销在保险业中的应用

1. 大数据与车险精准营销

1) 车险精准营销

保险公司为了提高其在车险市场中的竞争地位，需要通过精准营销将潜在的车险需求转化为车险产品的实际购买力。精准营销的实现离不开大数据技术的有效应用，因而保险公司要将大数据应用于其车险营销的全过程。

在车险精准营销的发展初期，保险公司要明确其车险精准营销流程和机构设置，并能够通过应用大数据技术设计出基于差异化定价的成套保险产品和服务，进而针对不同车险产品进行初步的宣传和推广。

在车险精准营销的发展中期，保险公司在大数据技术的应用下要对其车险产品进行进一步的细化，并利用大数据挖掘结果对其车险产品的研发过程和营销模式进行优化，不断扩充其车险产品和服务的宣传途径和营销手段。

在车险精准营销的发展后期，保险公司要以客户体验为核心目标，应用大数据挖掘的

结果对其差异化的车险产品进行再创新，对其车险产品的营销机制进行不断完善，进而使其客户满意度得到有效提升。

2) 相关案例

一直以来，平安财险都与百度搜索保持着良好的合作关系。当用户在百度搜索中搜索关键字“车险”时，平安财险的产品宣传就会出现在用户搜索结果页面中的显眼位置。除此之外，平安财险还利用大数据技术对其目标客户群体——车主的相关数据进行了重新梳理。

平安财险发现，在车主周围或远或近地聚集着汽车厂商、4S 店、汽车配件厂商、交通管理部门、加油站、导航服务提供商、保险公司等一系列组织机构，这些机构分别掌握着与车主以及投保车辆相关的各类数据。因此，它们在对车主进行研究时，突破了传统保险营销的局限性，从整个产业链的角度对车主的车险需求进行分析和判断。

平安财险进而从车主购车前、购车中、购车后的 3 个阶段出发，绘制出汽车生命周期的问题蓝图。该问题蓝图清晰地展示了车主在不同阶段所面临的不同问题和主要保险需求。例如，车主在购车阶段会考虑车险、购车贷款、经销商、车牌这几大类问题，而在每一大类问题下又会细分出更多的具体问题。平安财险基于其对车主在不同阶段的特征判断，为身处不同阶段的车主有针对性地推荐车险产品，使其车险产品的销售业绩得到了有效的提高。

2. 大数据与健康险营销

1) 健康险精准营销

健康险精准营销的思路及过程与车险精准营销并不存在太大的不同，都是基于大数据分析和挖掘的结果了解其客户偏好和保险需求，进而有针对性地进行营销，以实现营销效率的有效提高。这里不再对重复的内容进行赘述。

健康险精准营销中需要特别注意的问题在于对营销时机的把握。人们的健康管理是一项长期活动，但人们在没有患病恐惧时通常并不具有购买健康险的行为动机，而患病之后购买健康险也不再具有意义。因而健康险营销的最佳时机在于潜在客户具有患病恐惧之时，即发生医疗咨询行为之时。随着移动互联网技术的飞速发展，网上医疗咨询凭借其便捷性已成为人们进行简单医疗咨询的主要方式。保险公司通过利用大数据技术能够了解其潜在客户的健康状况和主要健康顾虑，进而向特定的潜在客户有针对性地推荐相应的健康险产品，从而实现健康险的精准营销。

2) 相关案例

法国 GMF 保险公司通过利用大数据技术对其 3 亿潜在客户的相关资料进行分析，建立了客户全生命周期的价值模型，进而使其获取新客户、进行交叉销售和追加销售的效率得到了极大的提高。在这一客户数据分析过程中，GMF 保险公司将其自身的客户数据库与第三方的客户数据和人口统计数据相结合，利用其建立的大数据分析平台对其所掌握的数据进行处理，并对其中的 1500 多个变量进行了不同角度的分析，进而从中找出了各种场景下保险产品销售与变量之间的相关关系，并在此基础上制定出具有针对性的营销推广策略。

国内的泰康人寿保险公司建立了语音记录的大数据分析平台，对其客户拨打 95522 的通话进行记录和分析，进而对这些客户进行了多样化的标签划分，如老年人、商务人士、大学生、运动员、医生、母亲、孕妇等。在其保险销售人员进行展业时，被展业客户的相关标签将在第一时间被销售人员获取，进而使销售人员能够以合适的销售方式向客户有针对性地推荐其保险产品。泰康人寿保险公司将此语音分析结构与其营销手段相结合，创造了千万元的保费收入。

5.4 欺诈识别

据统计，保险公司有 2/3 的支出被使用在理赔处理和赔款支付上，而作为伴生顽疾的保险欺诈行为仍有增无减、屡见不鲜。恶意保险欺诈的存在不仅严重损害了其他投保人所享有的正常权益，而且在一定程度上制约了保险服务社会的功能。随着科学技术的快速发展，保险公司通过大数据技术能够使其理赔处理、损失预防和欺诈识别的能力得到有效提高。

5.4.1 保险欺诈

1. 保险欺诈的主要表现形式

保险欺诈的主要表现形式有以下几种类型。

1) 虚构保险标的

即投保人就其在现实中并不存在的保险标的向保险公司投保，并在订立保险合同后谎报该保险标的发生盗取等保险事故，向保险公司骗取相应保险赔款的欺诈行为。该欺诈行为多发生于财产保险领域。

2) 不具有可保利益

即投保人就其不具有可保利益的保险标的向保险公司投保，并在保险合同订立后积极促成保险事故发生，进而骗取相应保险赔款的欺诈行为。该欺诈行为多发生于人寿保险领域。

3) 标的风险状况的虚假陈述

即投保人对其保险标的风险状况故意向保险公司隐瞒或做虚假告知，进而使保险公司在错误判断投保人风险状况的基础上对其进行承保，一旦达到风险条件投保人就可以从保险公司获取相应保险赔款的欺诈行为。该欺诈行为多发生于人寿保险领域。

4) 超额投保

即投保人凭借不实的相关单据以高于其保险标的实际价值的金额向保险公司进行投保，以期在风险事故发生后获得额外收益的欺诈行为。

5) 重复投保

即投保人就同一保险标的的同一保险利益在 2 个或 2 个以上保险公司进行投保，以期在风险事故发生后获得额外收益的欺诈行为。

6) 出险后投保

即投保人在特定保险事故发生后以不当手段对保险事故进行掩盖，进而向保险公司投保，以期就已发生的保险事故保险合同成立后获取相应保险赔款的欺诈行为。该欺诈行为多发生于人寿保险领域。

7) 主观故意出险

即投保人在保险合同成立后，带有主观故意性地促成保险事故发生的欺诈行为。

8) 虚假保险事故

即投保人的保险标的在保险期内并未发生保险事故，但投保人故意制造保险标的出险的假象，以期获得相应保险赔款的欺诈行为。

9) 夸大损失金额

即在保险标的出险后，被保险人不积极对损失进行合理的控制，甚至进一步加重损失，以期获得更多保险赔款的欺诈行为。

2. 保险业反欺诈工作中的问题

由于目前我国保险公司的反欺诈体系建设较为薄弱，使得我国保险欺诈现象非但没有得到遏制，反而有上升的势头。目前我国保险业反欺诈工作中所存在的问题主要表现为以下几个方面。

1) 对反欺诈工作的基础投入不足

国际保险监督官协会(IAIS)的经验数据显示，保险欺诈事件的赔付金额约占总赔付金额的 10%～20%。但在大多数保险公司中，专门从事保险反欺诈工作的人员数量不足全体员工数量的 10%，且在保险公司的基层分支机构中从事反欺诈工作的人员数量更少。

2) 未形成全国集中的反欺诈处理中心

虽然大多数保险公司都已在其内部建立起专门负责反欺诈工作的部门，但不同地区的分支机构之间、不同机构层级之间的反欺诈工作缺乏协调性，相关风险数据也并未在部门之间实现共享。由于诸多保险公司内部尚未形成全国集中的反欺诈处理中心，因而保险公司对于跨区域、跨机构的欺诈风险缺乏识别能力。

3) 大数据分析的思维方式缺失

虽然保险公司内部积累着海量的客户基础数据，但部分保险公司由于过分依赖传统的反欺诈工作方式且缺乏大数据思维，并没有利用大数据技术对这些价值数据进行有效的分析和挖掘，隐藏在这些海量数据中的欺诈线索也难以被发现。

4) 传统风险控制体系维度简单

在保险公司传统的风险控制体系中，对欺诈风险的排查大多是通过固定的程序化风险监测模型实现的。由于传统风险监测模型的维度简单且样本数量有限，致使模型与实际风险场景之间的匹配程度较低，监测欺诈风险的能力也极为有限。

5) 保险公司之间的数据共享机制缺失

由于我国保险市场的竞争较为激烈且行业内部尚未建立统一的信息共享平台，因而各保险公司都将其所掌握客户信息视为自己的核心资产而不愿与其他保险公司进行数据共享。这也使得同一主体的保险欺诈行为能够在不同的保险公司中重复发生。

6)　法律制裁不力、犯罪成本低

由于我国现行法律对保险诈骗行为的量刑较轻，有关部门对其在具体实践中的惩治职责认识不清，导致我国对保险欺诈行为界定不准、惩治力度有限，进而使有保险欺诈行为的不法分子得不到应有的量罪和制裁。而较低的犯罪成本和有限的法律制裁是导致保险欺诈事件频发的主要外部原因。

5.4.2　大数据与保险反欺诈

1. 大数据与欺诈识别

由于保险欺诈具有专业性和隐蔽性的特点，因而保险行业内部主张将大数据技术应用于欺诈识别工作当中的呼声日渐高涨，且有部分保险公司已经开始了大数据反欺诈的实践。

1)　大数据技术的优势

由于保险市场中的竞争日趋激烈，诸多保险公司都在努力提高自身的运营效率。就理赔运营环节来讲，保险公司需要在有效欺诈识别的基础上实现理赔流程的精简和理赔时效的提高。大数据技术在这一方面有很好的应用前景。

保险公司通过运用大数据技术对其所掌握的海量客户数据进行充分的分析和挖掘，能够从中找出对保险欺诈的发生影响最为显著的因素，以及这些影响因素的取值区间，进而构建出大数据保险欺诈识别模型。保险公司的理赔人员能够通过运用大数据保险欺诈识别模型对每个具体的理赔事件进行有效的欺诈风险评估，进而依据评分的高低对是否立即支付理赔金、是否进行实地勘查等问题做出决策。

随着科学技术的快速发展，保险行业中已出现针对理赔事件的智能勘查系统。智能勘查系统可以及时地为保险公司提供与保险标的出险状况相关的各项指标数据，进而帮助理赔人员从中找出异常状况并及时采取应对措施。

在利用大数据分析对欺诈风险进行监测的基础上，保险公司的理赔运营效率和客户体验能够得到有效提升。在大数据应用的背景下，保险公司能够对实时获取的保险标的出险信息进行快速分析，进而及时且主动地向其客户提供保险理赔服务。例如，客户在驾驶车辆的过程中发生保险事故，保险公司通过其在投保车辆上所安装的车载信息系统能够及时地获取出险报案信息，进而在客户提出理赔申请之前主动向客户提供理赔服务以及更多适宜的增值服务。

2)　大数据反欺诈流程

大数据反欺诈流程如图 5.10 所示。从图 5.10 中我们可以看出，反欺诈工作是从索赔人在保险标的出险后提出索赔申请(或由相关具有感知能力的信息系统发出实时警报代替索赔人提出申请)开始的。

保险公司在收到相关申请后将自动进入审核环节，即利用大数据技术对其所掌握的与投保人和保险标的相关的基础数据、由智能勘查系统及时反馈的与出险状况相关的实时数据进行处理和分析，对引起风险事件发生的主要因素进行识别和判断。在这一环节中，智能勘查系统能够向保险公司提供视觉化的信息并为其揭示潜在的犯罪网络，进而帮助保险公司对高风险索赔给予必要的关注。

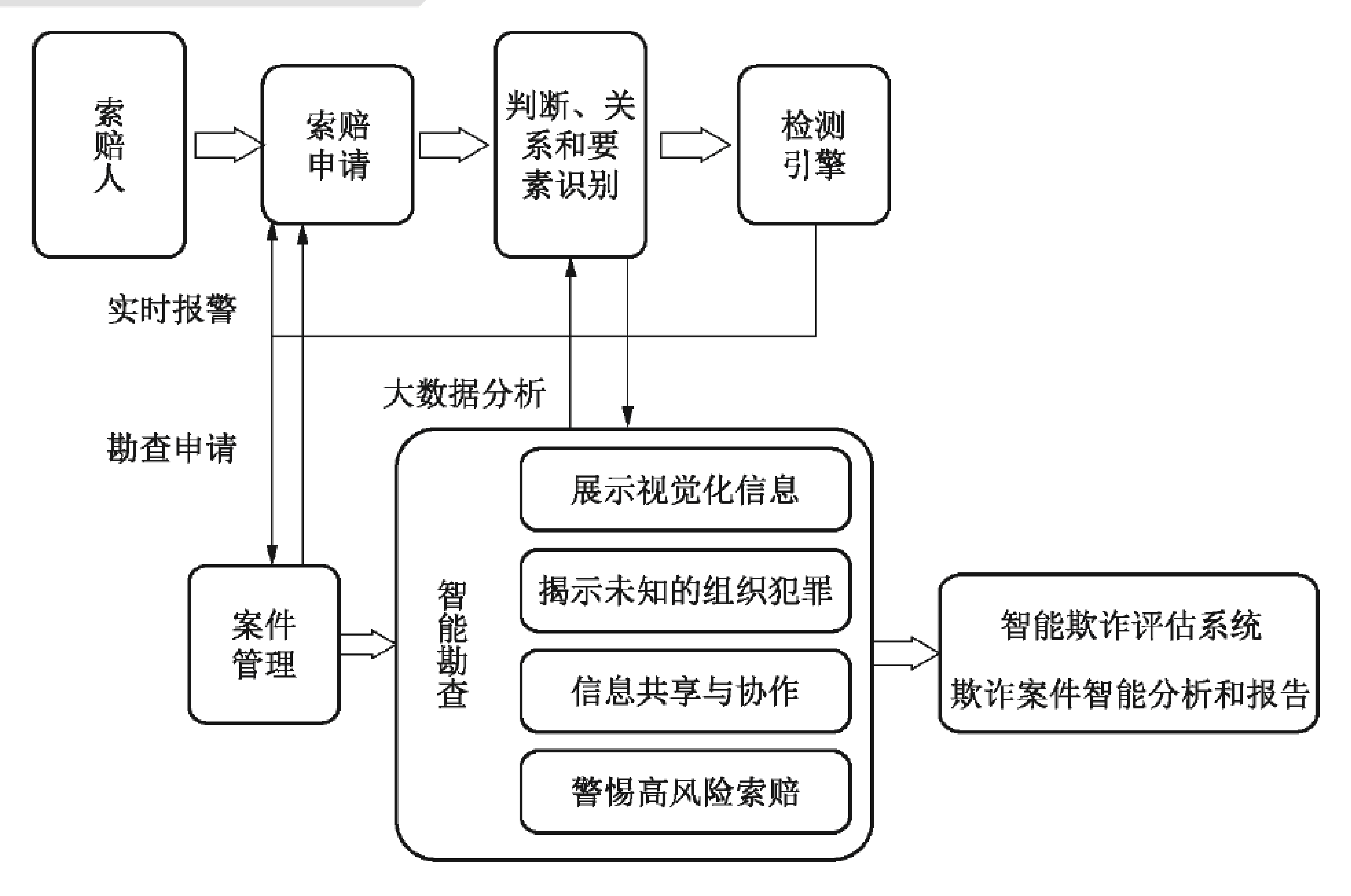


图 5.10　大数据反欺诈流程

接下来，要将上述大数据分析的结果接入智能欺诈评估系统，进而对该项理赔案件的欺诈风险进行评估：若该案件评分较高，则做出直接理赔的决策；若该案件评分较低，则做出进一步审核(如进行人工实地勘查)的决策。

借助大数据技术对海量数据的快速处理和分析能力，基于该反欺诈流程的欺诈识别工作十分高效：不仅使审核时间得到大幅缩短，而且是审核的准确性得到大幅提高。

数据越完整多样，基于大数据技术的反欺诈工作效率就越高，即数据资源的可靠和完整是大数据反欺诈工作高效进行的基础。因此，保险公司要对理赔历史记录、保单信息、医疗保险数据、事故统计数据、征信记录、犯罪记录、社交网络数据等相关数据信息进行有效的整合和存储。

【案例 5.4】大数据对保险反欺诈工作效率的提升

南非最大的短期保险产品供应商——Santam 保险公司也曾被保险欺诈所困扰。最初 Santam 保险公司为了应对可能存在的保险欺诈，放慢了其理赔处理速度——用至少 3 天的时间对理赔案件进行审核，这无疑使 Santam 保险公司良好的客户服务声誉受到严重影响。之后 Santam 保险公司开始采用基于大数据的欺诈风险分析和解决方案，使其欺诈识别能力得到大幅提高。在该系统中，Santam 保险公司依据其已经确定的风险因素对每个理赔案件进行评估，并根据理赔案件风险程度的不同采取不同处理方式。Santam 保险公司借助该大数据欺诈识别系统不仅节省了数百万美元的保险欺诈损失，而且还使其低风险理赔案件的处理时效得到有效提升，绝大多数正常的理赔案件能够在不到 1 个小时的时间内处理完成。

美国 Allstate 保险公司利用大数据技术分析出保险欺诈的潜在规律，进而使其理赔欺

诈的损失得到大幅降低。Allstate 保险公司借助大数据技术对理赔数据、投保人数据、相关网络数据和揭发者数据进行有效的整合和挖掘，建立起大数据欺诈识别系统；进而将所有理赔请求首先接入到该大数据欺诈识别系统之中进行自动处理，然后再将可疑的理赔请求交由特别调查部门进行人工审阅。Allstate 保险公司通过利用大数据技术成功将其保险欺诈发生率降低 30%，将其欺诈识别准确率提高 50%，并将其理赔成本节约近 3%。

(资料来源：甘肃信息网)

2. 大数据反欺诈工作的重点

1) 对相关数据进行有效利用

大数据时代背景下，新信息技术的出现和应用为保险公司的反欺诈工作提供了更多的可能。保险公司大数据反欺诈工作的核心就是对相关数据进行有效利用。

(1) 建立信息共享机制。

数据是保险公司进行反欺诈审查的基础。因而保险公司为解决信息不对称问题，要利用大数据技术建立信息共享平台，为其进行反欺诈工作奠定良好的基础。

(2) 管理和整合相关数据。

在对内部数据进行整合的基础上，保险公司还要积极地与第三方合作以获取更多与投保人和保险标的相关的数据，进而形成对业务风险更准确的判断。保险公司要对其所掌握的相关数据进行有效的管理和整合，以在保护个人隐私的基础上实现数据价值的充分利用。

(3) 可视化关联分析技术。

保险公司利用大数据技术对海量信息数据进行专业化处理，能够以直观的方式将承保、理赔、客户等相关层面的数据中所隐含的信息表现出来，进而使保险公司反欺诈工作的脉络更加清晰和明确。

(4) 对信息进行量化分析。

在获取数据并对数据进行整合的基础上，保险公司通过对相关数据进行量化分析，能够借助预测技术建立用于欺诈识别的统计分析模型。在将各例理赔案件的相关数据接入该欺诈识别模型中后，保险公司就可以根据模型给出的评分来判断各例理赔案件中的欺诈风险。数据在反欺诈工作中的应用流程如图 5.11 所示。

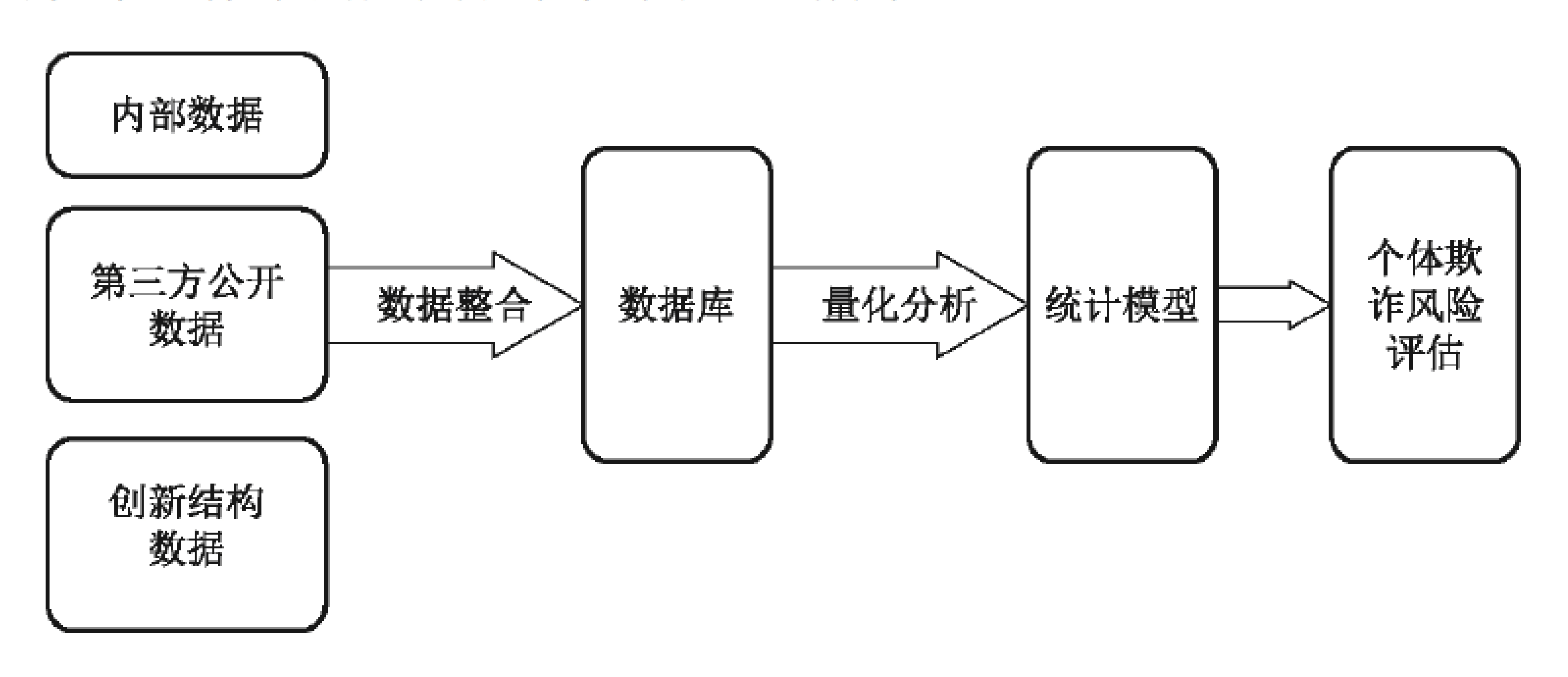


图 5.11　数据在反欺诈工作中的应用流程

2) 建立科学的承保和理赔规程

为有效地控制欺诈风险，保险公司要对其承保和理赔环节的工作机制和流程进行优化，将反欺诈工作的重心从被动的事后控制转移到主动的事前控制之上。

(1) 承保环节保证质量。

为从源头上遏制保险欺诈的发生，保险公司要保证其承保环节工作的高质量。保险公司可以利用大数据技术量化分析投保人的投保动机，进而在订立保险合同前实现对投保人的欺诈风险评估。保险公司还要建立有效的承保审核制度、信息沟通制度和岗位考评制度，为高质量承保的实现提供制度保证。

(2) 建立两级勘查制度。

即在对理赔申请进行审核时，对于欺诈风险评估模型给出较低评分的理赔申请要进行实地勘查，并在实地勘查过程中严保查勘质量。对于一些特殊的理赔申请，保险公司还应当通过复勘来提高其在审查环节中的工作质量。

(3) 建立规范的理赔制度。

保险公司要建立接案人、定损人、理算人、审核人和审批人之间的分离制度以及实地勘查人员之间的制约制度，并对相关风险评估数据和实地勘查报告进行有效的存储和备份。此外，保险公司还要建立严格的追责制度，一旦发生人员违规问题必须严肃处理。

3) 强化行业内部协作

(1) 全面推进行业信息共享。

为获取更多的客户信息，保险公司可以在保证客户隐私和相关数据安全的前提下在行业内部建立统一的信息共享平台，以打破各保险公司之间的数据孤岛。将分散在各保险机构的相关数据按类型的不同进行分类存储和有限共享，进而使共享数据在保险反欺诈工作中的内在价值被充分释放。

(2) 制定行业大数据规划。

有关部门要结合大数据的时代背景对与保险反欺诈工作相关的法律法规进行完善，进而为保险公司大数据反欺诈提供良好的政策环境。同时保险业协会要对行业数据标准进行改进和补充，以保证行业共享数据的质量。此外，还要建立行业数据的分析模型和研究框架，并建设与行业大数据相配套的数据安全防护体系。

4) 推进保险业信用体系建设

要加快建立保险行业内部的统一信用平台，对投保客户以及从业人员的信用状况进行记录和评价。基于“失信惩戒”的原则，有关部门要在行业内部建立起行业黑名单制度以及市场退出机制，以实现对失信主体的有效约束和惩戒。

5.4.3 大数据与车险反欺诈

1. 我国车险反欺诈工作现状

1) 市场整体环境层面

近年来，我国车险市场中频发的欺诈现象严重阻碍了车险市场的有序、健康发展。为

保护保险消费者合法权益，切实防范和化解保险欺诈风险，中国保监会于 2012 年 8 月出台了《关于加强反保险欺诈工作的指导意见》。该意见指出，各保险公司和保险中介机构内部应针对欺诈风险建立反欺诈制度机制。

随着打击力度的不断加大，车险欺诈行为日益呈现出多样且隐蔽的特点，传统的车险欺诈识别方法已难以对相关欺诈风险进行防范。大数据技术的出现为车险反欺诈工作提供了更多的可能，保险公司能够借助大数据技术从海量数据中识别出潜在的车险欺诈行为，并有针对性地对其进行防范。

2) 行业内部工作层面

从保险公司具体的经营管理层面来看，目前我国保险公司采用的车险反欺诈工作方法主要有：独立调查人、内部调查人、相关费用审核、集中定损、定损复核、理赔审核、数据分析平台等相关反欺诈方法。其中，集中定损是保险公司为防范欺诈风险对其业务流程进行优化的表现，主要包括快速定损和拆检定损两种方式，其具体的业务流程如图 5.12 所示。

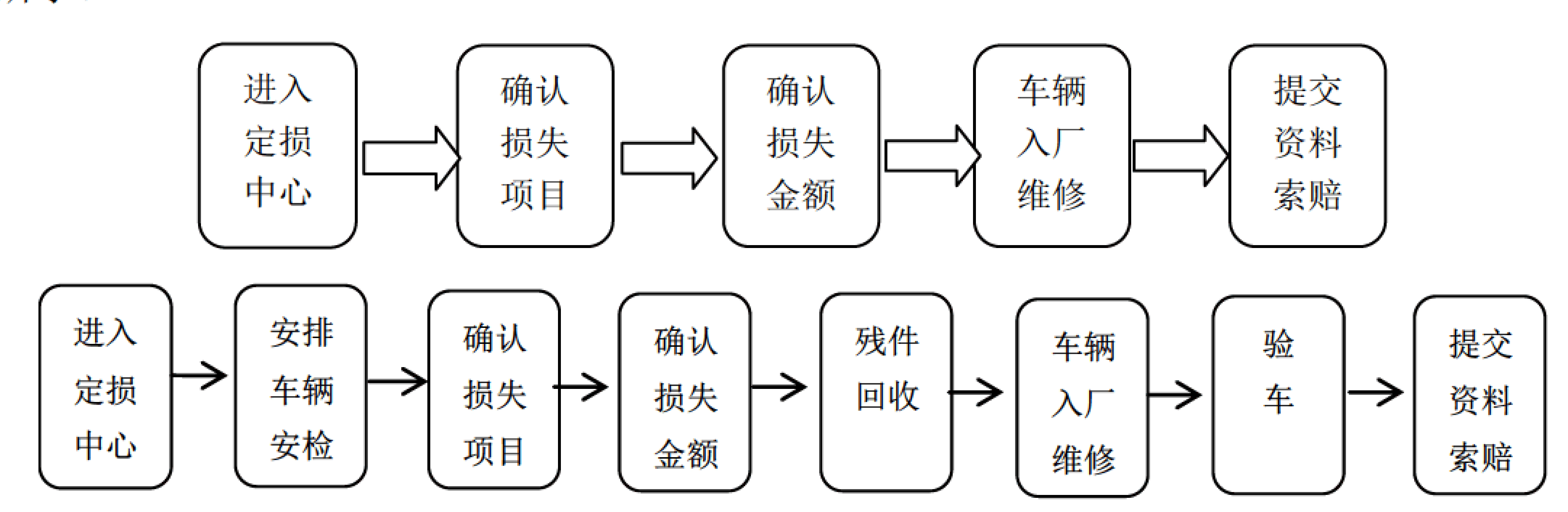


图 5.12 保险公司集中定损的业务流程

虽然目前已有个别保险公司建立了相对完善的车险反欺诈机制，但大多数保险公司的车险反欺诈工作仍未实现专业化的管理和运作。随着大数据时代的到来，许多保险公司都在借助大数据技术努力提高自身的车险反欺诈能力，并取得了一定的成效。因此，我们可以预见在不远的将来，保险公司的车险反欺诈工作将会实现质的飞跃。

2. 车险欺诈的风险识别因子

我国的车险主要有机动车交通事故责任强制保险(即“交强险”)和商业车险两种类型，其中商业车险又可分为基本险和附加险两种类型。从车险涉及主体和车险赔付过程两个方面来看，车险欺诈的风险识别因子包括但不限于如图 5.13 所示的 30 个识别因子。

在图 5.13 所示的风险识别因子涵盖了与投保人、驾驶员、保险公司从业人员、投保车辆、投保车辆的维修厂商以及保险中介机构相关且包含一定欺诈风险信息的数据指标，以及出现在车险经营过程(包括投保、出险、勘查和理赔环节)中的包含一定欺诈风险信息的数据指标。以这些指标为基础，能够建立起有效的车险欺诈风险识别模型。

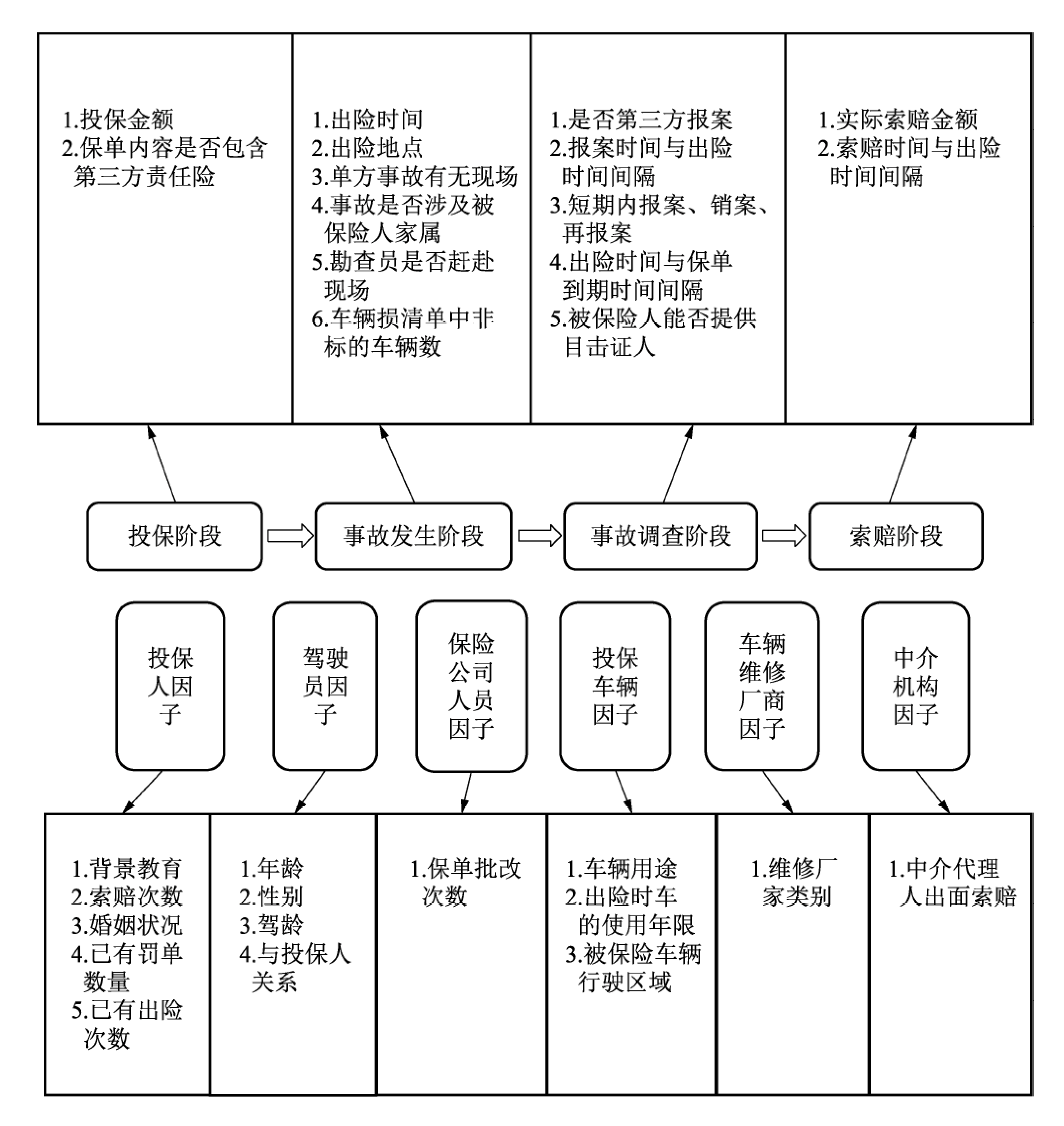


图 5.13 车险欺诈的风险识别因子

3. 车险欺诈识别的理论模型

在大数据的时代背景下，基于海量数据所建立起的车险欺诈识别系统能够对车险欺诈风险进行有效的识别和防范。因而保险公司为提高其在车险反欺诈工作中的能力和效率，要在其内部建立起完善的车险欺诈识别系统，以对其所掌握的信息数据进行充分利用。基于前文所述车险欺诈的风险识别因子，可以建立车险欺诈识别的理论模型，如图 5.14 所示。

4. 车险反欺诈防范对策

1) 构建跨行业的客户信用数据库

各保险公司内部都存储着一定的客户信用数据，但各保险公司之间由于缺乏信息共享机制，其对客户信用状况的评估和把握仍不够准确。此外，其他行业中也存储着诸多与车

险欺诈行为具有相关性的数据，对这些外部数据加以利用能够有效地提高保险公司对车险欺诈的识别和防范能力。因此，要在保护客户隐私安全的基础上，建立涵盖相关内外部数据且能够在行业内部实现共享的客户信用数据库。

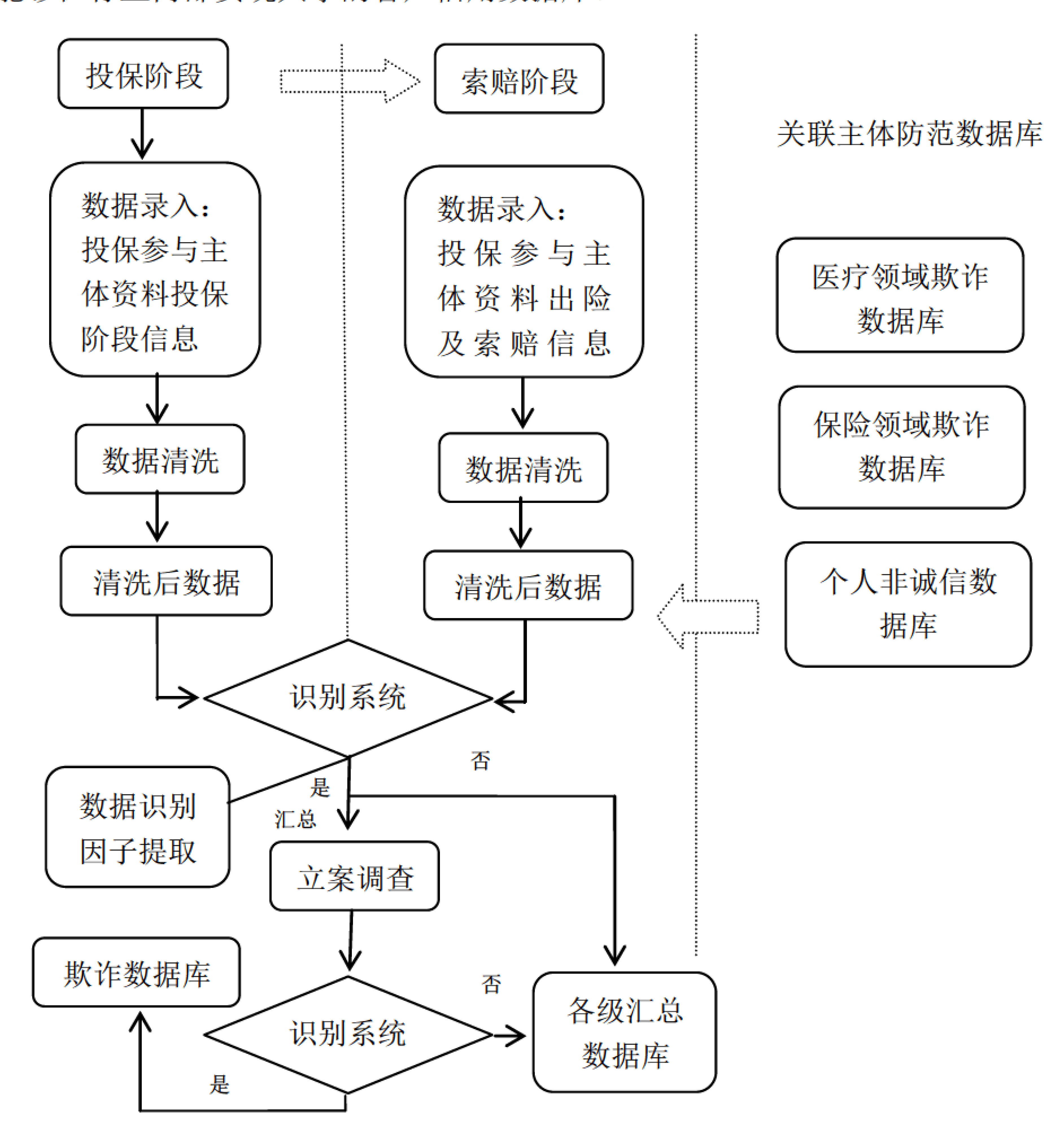


图 5.14　车险欺诈识别的理论模型

2)　推广使用车载信息系统

由于实时数据能够帮助车险欺诈识别工作实现动态调整，保险公司在对车险欺诈风险进行识别时，除了要利用内部存量数据和相关外部数据，还应当考虑借助车载信息系统获取与投保车辆相关的实时数据并加以利用。因而保险公司在进行车险承保时，要以合理的方式和手段鼓励其车险客户安装和使用车载信息系统。

3)　建立全国性的车险反欺诈联动机制

为更有效地开展车险反欺诈工作，要在保险业监管部门和自律组织、保险公司、交通管理部门、相关科研机构等相关主体之间建立全国性的车险反欺诈联动机制，从而实现对车险欺诈风险的多角度识别和全方位防范。

5.4.4 大数据与健康险的理赔风险

1. 健康险中的理赔风险

1) 健康险的发展困境

近些年来，我国的商业健康险在高速发展。根据保监会的统计数据显示，我国 2016 年前三季度商业健康险的原保险保费收入为 3430.41 亿元，同比增长 86.77%。虽然我国商业健康险有着良好的发展态势，但从事商业健康险业务的保险公司大多仍未实现盈利。究其原因，主要有以下两个方面。

(1) 目前保险行业内部对商业健康险的市场定位尚待进一步明确，从事商业健康险业务的保险公司尚未走上专业化的发展道路。

(2) 保险公司与医院等医疗机构之间存在信息不对称，进而导致其难以对医疗费用的赔付风险进行控制。相关赔付成本的居高不下正是阻碍商业健康险良好发展的突出问题。

2) 健康险中的理赔风险

商业健康险中的理赔风险主要包括客户的欺诈风险和医疗机构的过度医疗风险。其中，医疗机构的过度医疗风险最为突出。

客户的欺诈风险在商业健康险中并不是主要风险，但一旦出现便涉及较大的金额。商业健康险中的客户欺诈行为与其他险种中的客户欺诈行为在本质上并不存在较大的不同。因此，商业健康险中的反欺诈工作在内容和流程上也与其他险种类似。在大数据应用的背景下，通过对多类型的海量客户数据进行充分的分析和挖掘，保险公司能够用模型来刻画商业健康险中客户欺诈行为，进而高效地对每一例理赔案件进行审查并快速做出适当的行为决策。

医疗机构的过度医疗风险作为商业健康险中的主要风险，是由保险公司与相关医疗机构之间的信息不对称所造成的。过度医疗行为包括但不限于：药品用量超标、用药与患者实际医疗需求不匹配、医疗服务的非合理收费、药品的非合理定价、基于保障方案的非必要医疗行为等(见图 5.15)。相关医疗机构的过度医疗行为导致保险公司对其所承保的相关医疗费用负担着极高的赔付成本。据公开信息显示，仅由药品用量超标和非必要医疗行为两项所导致的保费资源浪费就达到了 20%～30%，再加上药品非合理定价、医疗服务的非合理收费等其他过度医疗行为的影响，保险资源的浪费比例高达 50%以上。因此，经营商业健康险业务的保险公司要想在该业务中实现盈利，就必须对其所面临的过度医疗风险进行合理且有效的控制。而医疗信息数据正是对过度医疗风险进行有效控制的关键。

2. 大数据与健康险的理赔风险控制

大数据时代的到来，为保险公司对其商业健康险业务中的相关理赔风险进行有效控制提供了数据基础和实现途径。

1) 与医疗大数据相结合

商业健康险的核心是医疗服务。随着我国的医疗信息化建设的逐渐深入和医疗数据库体系的不断完善，与具体医疗服务相关的数据资源将会被有效地获取和整合，进而使保险公司与客户和医疗机构之间的信息不对称问题得以解决。因此，保险公司要把握时机，积

极向有关部门争取接入医疗数据库的机会。此外，鉴于医院是医疗大数据获取的关键卡位，保险公司还应加强与医院之间合作，进而实现对客户健康状况和医疗行为的精准把握。总而言之，保险公司应主动顺应大数据的时代潮流，尽可能多地开发数据获取渠道，以提高其风险识别的准确性。

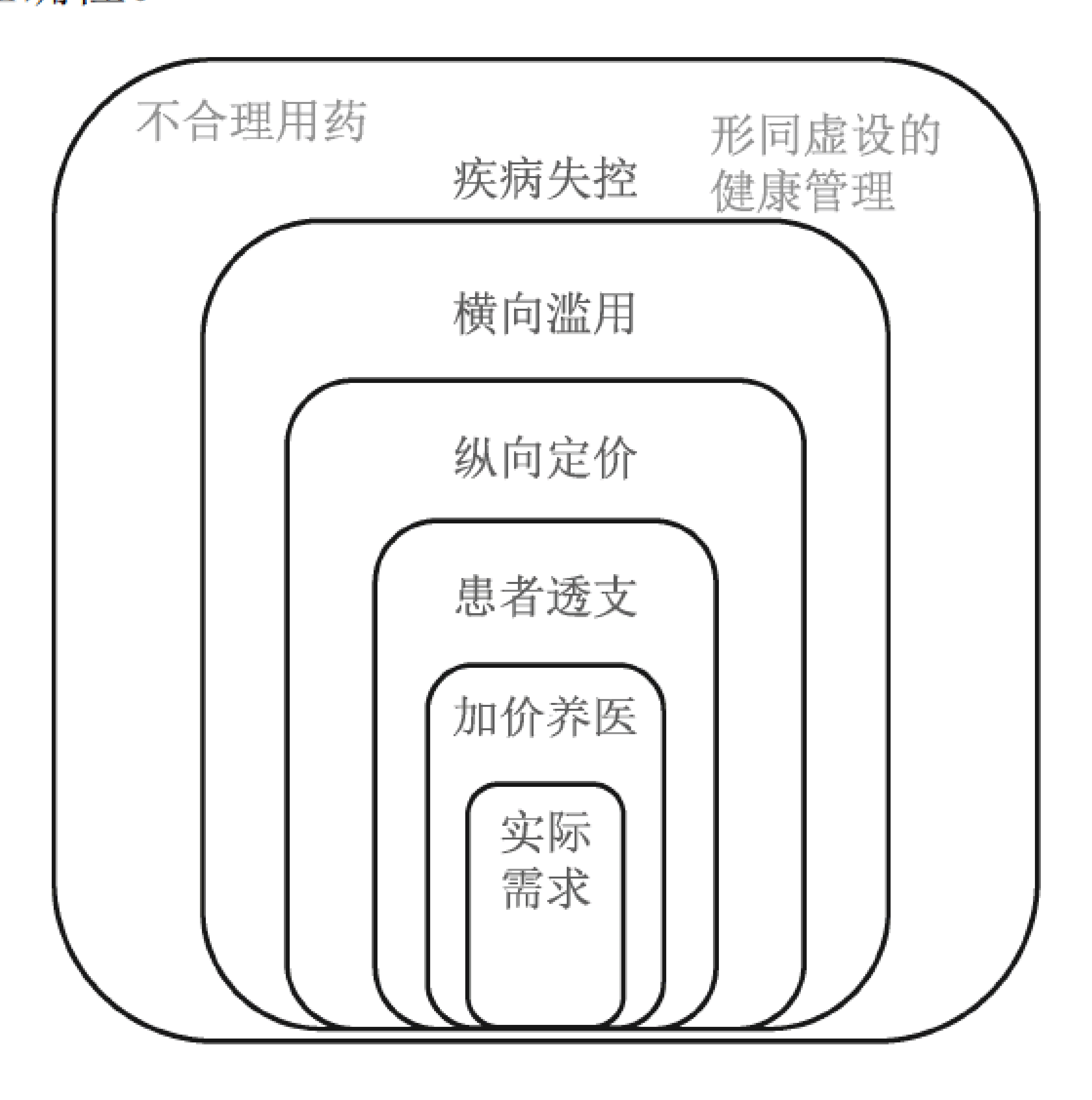


图 5.15　过度医疗行为

2)　合理评估医疗费用和质量

在获取海量医疗数据的基础上，保险公司可以利用大数据分析技术对相关医疗行为的费用和质量做出科学合理的评估。

由于具体的医疗服务行为难以被标准化，因而保险公司难以对医疗费用的合理性做出准确的评估。例如在心脏支架手术中，进行哪些方面的化验检查、采用何种麻醉方式、使用哪种心脏支架、支架的放置数量、术后需要多久的康复期、康复期内需要接受哪些化验检查等问题，都会因患者的身体状况和经济能力的不同而存在差异。结合医疗服务行为的这一特点，保险公司可以借助大数据技术找出同一疾病相关诊疗项目与用药情况之间的相关性，以专业的分组方法对相关诊疗费用的标准进行评估。

而对医疗质量的评估，保险公司可以从医疗过程评估和医疗结果评估两个方面进行。保险公司可以通过利用大数据技术对海量的医疗临床数据进行分析和挖掘，进而准确地判断出在不同疾病的诊疗过程中哪些医疗行为是必需的、哪些医疗行为是不合理的、所用药物是否是合理的、用药剂量是否是合理的等，即实现对医疗过程的评估。保险公司还可以通过利用大数据技术对海量的康复期数据进行分析和挖掘，进而对术后不良事件发生率、疾病复发率等相关指标进行判断，即实现对医疗结果的评估。

3)　大数据与健康险风险管理

如图 5.16 所示，通过利用大数据技术对以病历为中心的相关医疗数据进行挖掘，保险公司能够基于不同患者的具体健康体征和主要症状，对诊疗过程中所发生的相关医疗行为进行有效核查，进而实现对赔付金额的合理控制。

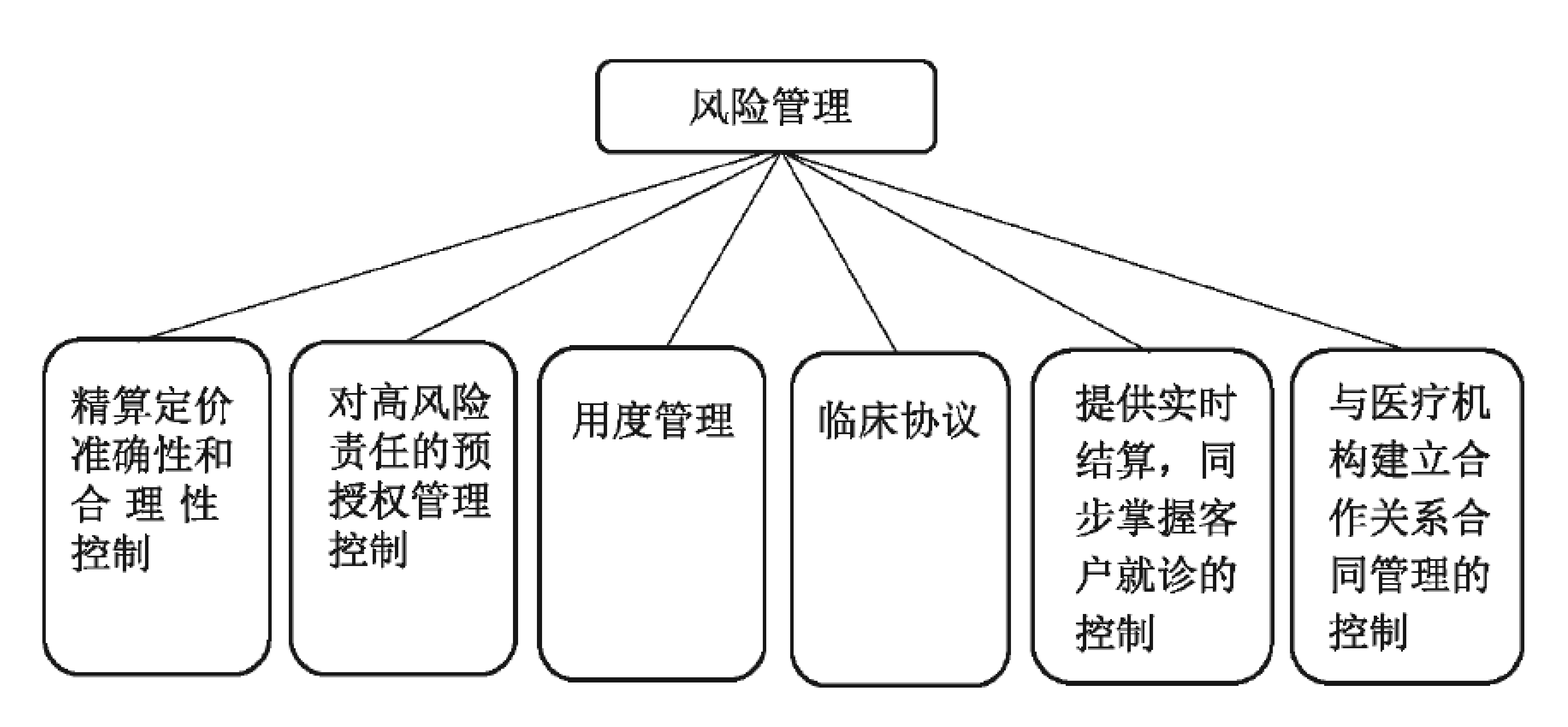


图 5.16 大数据与健康险风险管理

目前国内已有部分保险公司开始了在健康险中应用大数据技术的实践。例如，太平洋保险集团旗下的太保安联健康保险公司通过与阿里健康进行合作，将阿里健康所掌握的海量数据、风险控制引擎和人脸识别防伪等技术接入其理赔环节，使其控费能力得到有效的提升。

本章总结

- 大数据保险是指保险公司通过利用大数据技术对风险数据进行分析、处理和挖掘，使风险数据实现有效的价值变现。在此基础上保险公司通过其治理端和商业端的协同创新，使传统的保险服务方式和资源配置方式得以优化，从而实现保险产品、保险服务和保险业务模式的创新，进而更好地满足其客户需求并提供更为优质的保险服务。
- 在金融领域中，保险行业应用大数据相对较晚，应用水平也落后于银行业和证券业。这是因为银行业与证券业的数据服务平台建设较早，从而为大数据技术的应用奠定了良好的基础，而保险业的数据服务平台建设则相对较晚。而就保险业自身的大数据应用阶段而言，目前尚且处于大数据应用的初级阶段，即内部循环阶段。因而接下来保险业需要通过合理利用其内部数据并引入更多的外部数据来拓展大数据分析在本行业中的应用领域。
- 保险公司的承保定价能力是其在同业竞争中的核心竞争力。在大数据技术的应用下，保险公司过去的样本精算将升级为全量精算，风险定价模式将发生很大的改变。通过应用大数据技术，传统的保险精算中将引入更多的定价因素，保险公司能够根据客户的特定风险来调整承保定价，不仅能够使客户的差异化需求得到满足，还能使保险公司的承保风险得到降低，从而达到客户和保险公司双方共赢的目的。
- 保险公司在大数据背景下对客户保险需求进行分析时，相关的数据来源不再仅限于保险行业内部的保险客户数据，还包括保险体系以外的数据，包括与保险公司

进行合作的第三方自身所积累的客户数据、网络社交平台所积累的客户数据等。在对客户信息和客户行为的相关数据进行深度分析的基础上，保险公司能有效地预测客户需求、挖掘潜在客户并向其推荐适当的保险产品，从而实现个性化的精准营销。

- 保险公司通过运用大数据技术对其所掌握的海量客户数据进行充分的分析和挖掘，能够从中找出对保险欺诈的发生影响最为显著的因素，以及这些影响因素的取值区间，进而构建出大数据保险欺诈识别模型。保险公司的理赔人员能够通过运用大数据保险欺诈识别模型对每个具体的理赔事件进行有效的欺诈风险评估，进而依据评分的高低对是否立即支付理赔金、是否进行实地勘查等问题做出决策。

本章作业

1. 简述大数据保险的概念、特征应用阶段。
2. 大数据在保险行业中有哪些作用？
3. 大数据背景下的数据服务架构与传统数据服务架构有哪些区别？
4. 大数据是如何帮助保险公司实现承保定价能力提升的？
5. 简述基于 OBD+UBI 的车险费率厘定方式。
6. 大数据时代产生了哪些保险新营销方式？保险公司又该如何提高其精准营销能力？
7. 什么是垂直平台生态圈？构建的动因有哪些？
8. 大数据是如何帮助车险和健康险实现精准营销的？
9. 保险欺诈有哪些形式？保险公司如何利用大数据开展保险反欺诈工作？
10. 阐述大数据在车险和健康险领域中是如何帮助保险公司进行理赔风险控制的。

第 6 章

互联网金融中的大数据应用

本章目标

- 了解如何用大数据技术对第三方支付中的欺诈风险进行风险防范
- 掌握大数据技术在网络借贷中的应用
- 掌握大数据技术在互联网供应链金融中的应用
- 熟悉大数据在互联网消费金融中的应用

本章简介

互联网金融是金融行业的后起之秀，它的出现打破了传统金融行业的格局，也正渐渐地改变着人们的生活方式。它以互联网技术为依托，将金融服务线上化，推动了金融产品创新，但是由此产生的风险也是不可忽视的。而大数据技术的应用很好地解决了这个问题，同时也解释了互联网金融为什么能受大众欢迎而传统金融机构失宠的问题。本章从第三方支付、网络借贷、互联网供应链金融、互联网消费金融等方面着手，通过各个典型的案例介绍互联网金融中的大数据应用。



6.1 基于大数据的第三方支付欺诈风险管理

6.1.1 第三方支付中的欺诈风险

欺诈风险是第三方支付机构面临的主要外部风险。由于第三方支付依托于互联网和电子商务，而目前我国存在较为严重的网络漏洞，信息安全没有得到有效的防护，这给不法分子带来了犯罪的空间。在这种虚拟环境下，不法分子更容易伪装自己的身份进行交易，欺骗消费者。

基于第三方支付的欺诈行为主要包括以下两种形式。

第一种，不法分子通过木马病毒等方式在消费者不知情的情况下，侵入消费者的第三方支付客户端盗取相关信息，从而实现资金盗取，产生欺诈风险。这种诈骗方式需要以一定的计算机技术为支撑，但诈骗行为一旦发生，消费者就很难在事前察觉，也很难在事后挽回损失。

第二种，不法分子的欺诈行为是利用消费者自身的防骗意识较弱得以实现的。一方面，不法分子通过注册一家网店，然后推出一些优惠活动吸引消费者参与，并告知若想参与此活动只能通过打开所给链接或者扫描所给二维码进行支付，从而诱使消费者付款。另外一方面，不法分子是以第三方支付机构的名义给消费者发邮件或信息，以用户的账号密码不安全或者补充个人信息等为由，通过窃取消费者的账户信息来实现其欺诈行为。

目前随着网络交易的丰富，各式各样的欺诈形式层出不穷，其本身大都与第三方支付机构无关，但是不法分子正是利用消费者对第三方支付机构的信任或第三方支付机构本身运作时存在的漏洞进行不法行为，最终使消费者蒙受损失。

欺诈风险存在的本身不是第三方支付机构违规行为导致的，但欺诈风险的蔓延会打击消费者的信心，严重危及第三方支付行业的市场形象。现阶段，国内的第三方支付机构对此类事件的发生都设有相关的免责条款，以支付宝为例，其明文规定："本公司对您所交易的标的物不提供任何形式的鉴定、证明的服务。"这意味着它本身不承担相关的监督责任。此外，由于对第三方支付机构业务操作的具体流程没有相关规定，导致无法对其注册用户的信息进行有效的核实和管理，这也使得不法分子能够利用虚假信息来实现网络欺诈。

但是基于电子商务的虚拟性、复杂性，第三方支付机构有责任采取更为有效的措施，包括风险识别、安全认证、建立健全垫付与追偿制度等，防范和化解欺诈风险，切实维护消费者的权益和第三方支付机构的信誉和安全。欺诈风险的发生不仅会影响消费者的交易， 也会破坏健康有序的交易秩序，对第三方支付机构本身也有很大的影响。因此，欺诈风险也是第三方支付平台在运营过程中需要解决的一个难题。

6.1.2 大数据应用与欺诈风险防范

对第三方支付平台而言，大数据是它浑然天成的优势。一方面，第三方支付涉及资金交易，它在客户注册使用时便可采集到客户的基本个人资料，如个人信息、身份证信息、

银行卡信息、财产信息等。这些数据通常被认为是非常有价值而且是较难获取的。另一方面，第三方支付在十几年的发展过程中，不断积累客户的海量历史支付信息本身就是大数据。这些大数据具有体量大、覆盖全、质量高的特点。第三方平台完全可以利用好已有的大数据，从而进行大数据风控，防范欺诈风险。

具体而言，第三方支付公司运用计算机技术，建设一个云端的动态数据库，数据库中储存和记录着客户的基本个人信息和交易信息。之后通过已有的数据进行科学的管理、合理的分类，并通过一定的算法建立风险控制模型。

大数据风控更侧重云端实时风险分析，通过对用户行为数据的关联分析发现蛛丝马迹，从而阻止欺诈的进一步发生。它的亮点在于，即便客户已经处于不安全状态，比如用户因木马钓鱼等原因导致账户密码等信息已发生泄露，经过云端的数据关联分析也能判断账户是否异常，并立即做出反馈。

大数据技术对第三方支付欺诈风险防范的应用，主要从以下 4 个场景展开分析。

1. 注册场景

注册场景中主要面临垃圾注册的欺诈风险。详细地说，就是欺诈者可能会在某一个第三方支付平台上注册很多账号，而这些账号通常不会有实际的交易，是一堆“空号”。欺诈者这么做可能出于两种目的。第一，第三方支付平台可能通过营销活动吸引新注册用户，如注册送红包、优惠券、礼品等(大多以红包为主)。而欺诈者正是利用这种活动，通过注册多个账户“聚沙成塔”，以此获利。第二，欺诈者通过注册多个账户，很有可能是为后续的洗钱、盗卡、诈骗等欺诈行为埋下伏笔。

在注册场景下，存在的欺诈风险有以下两个特征。

(1) 一般地，互联网企业都会在用户注册界面设置图片验证码，以对注册者进行“人机图灵测试”，即判断注册者到底是人还是机器，以防止恶意的注册行为。然而这种方式也存在着一定的漏洞。因为目前市场上已经存在这样一个黑色产业链，欺诈者通过雇佣劳动力，让受雇人进行有偿的识别验证码的工作。在这种情况下，图片验证码显然形同虚设。

(2) 当用户进行第三方支付的账号注册时，一般需要进行手机号验证并绑定，旨在限制单个客户注册的账号数，防止一人多号的现象。而市场上存在着专门设定虚拟手机号码以进行验证的收码平台。欺诈者完全可以通过该平台以廉价的方式获取非常多的虚假手机号进行注册，且通过收码平台接收提供虚拟账号短信验证码，欺诈者可以轻松绕过手机短信验证码环节。

针对上述欺诈风险，第三方支付平台可以运用大数据技术，利用云端的数据库，分析用户注册行为是否存在异常。例如，看注册者注册来源请求的 IP 地址是否是代理、同一个设备上发起的注册行为是否过于频繁；此外，平台可以通过外部的或者自有虚假手机号码数据库进行识别，并建立一个定期的清洗机制，确保数据的精准性。

2. 登录场景

在第三方支付平台的登录场景中，主要面临账户盗用以及撞库的欺诈风险。

首先，对于账户盗用风险，往往用户是因木马钓鱼或互联网泄露数据等各种不安全操

作，导致持有的账号密码信息被盗取。欺诈者在获取账号密码信息后，会尝试越权登录访问用户支付后台页面，进而发起盗卡交易行为，如购置虚拟商品等，导致用户账户资金损失。此外，越权访问获取支付账户的个人信息利用价值非常大，往往也会被欺诈者在黑产市场中反复交易利用。

对于账户撞库风险，更是第三方支付平台乃至所有互联网企业头疼的一个难题。由于互联网时期是信息爆炸的时代，许多用户会在网络上注册开通许许多多的应用或网站服务。许多人为了方便，会在不同的网站和应用中注册同一个账号密码，这就给客户的信息安全及财产安全带来了不安全因素。只要其中某一个应用或网站的安全性较弱导致被黑客攻陷，那么该网站的账号密码数据库就会发生泄露，这个过程称为“拖库”。黑客在完成“拖库”之后，会对数据进行清洗、封装，并对一些有价值的平台进行定向撞库攻击，即用已泄露的账号密码进行模拟登录尝试，若尝试成功即意味着单个账户撞库成功。账户撞库风险的危害在于导致数据大规模泄露，同时黑客攻击的成本正随着工具自动化而逐步降低。

面对这种登录场景中出现的欺诈风险，可以利用大数据技术采取以下几种风险防范措施。

(1) 判断用户的登录行为异常。一般情况下，用户在短时间内通常只会在同一个 IP 地址进行连续登录，而短时间内在不同的 IP 地址登录的概率很低。当出现用户在极短时间内连续登录且每次登录的 IP 解析位置距离偏移过大时，这很可能是欺诈者在挂 IP 代理进行登录，意图隐匿登录来源。为此，第三方支付平台可以运用规则模型对登录用户的登录时间间隔和 IP 解析地址偏移进行测算，当检测到上述异常行为时，那么系统可以对此用户加大关注度。

(2) 判断用户登录环境异常。在撞库过程中，黑客往往会使用成熟的工具程序进行批量模拟登录接口。那么，我们可以在登录页面布控人机识别检测程序，判断登录来源设备是否缺失、伪造，用户交互的行为是否存在缺陷。

(3) 判断用户登录习惯异常。一般情况下，第三方支付平台用户的账户常用设备、常用登录地都是稳定的，而出现登录习惯异常时，很有可能是出现了账户被盗用的情况。对此，第三方支付平台可以运用大数据技术，对用户登录行为进行长时间的跟踪分析，分析出账户常用设备、常用登录地等行为习惯。在此数据分析基础上，建立一套可信设备体系，即对于在可信设备上的行为业务应快速通过放行，而发生在非可信设备上的行为应加入重点关注。

3. 绑卡场景

在用户绑卡场景中，第三方支付平台通常已经从卡的维度进行风险防控，主要是对卡进行卡号、身份证、姓名、预留手机四要素进行验证。然而这种风险防控仍然存在一定的漏洞，并不能完全管控欺诈风险。

对欺诈者而言，首先会通过制作手机木马钓鱼软件进行传播，主要以手机端为主要的传播渠道。一旦有用户的手机不幸中招，那么该手机接收的校验短信码会被木马钓鱼软件拦截控制。欺诈者通过该方式能够收集一批受控制的手机及号码，随后从黑市交易的泄露

数据中进行筛选匹配，找出匹配的泄露的银行卡、姓名、身份证信息。欺诈者可以利用这些信息在其他第三方支付平台注册新账户，并以客户的身份完成绑卡操作，之后再将资金转走，也可以用此修改账户的密码。

面对绑卡场景中的欺诈风险，可以利用大数据技术对绑卡用户的信息、设备、IP 等维度进行关联分析，对于中介或者团伙的批量绑卡行为特征进行快速甄别。若出现异常行为则立即进行反馈，如冻结账号、通知真实用户等。

4. 支付场景

在支付场景中，主要面临的欺诈风险是盗卡支付及监管层面要求的反洗钱反套现监控。

盗卡支付风险还是源于个人隐私信息泄露，用户在绑卡的时候，其银行卡、身份证、手机号信息很有可能被黑客获取并将信息打包转卖。现在也有不少欺诈者通过各种渠道如邮件或伪基站，发送钓鱼链接诱导用户主动送上自己的信息。被盗客户账户的交易行为会出现异常，例如原本交易量较小的账户突然发生连续的多笔支付操作，或者银行卡出现莫名的支付行为。而洗钱套现行为更多是金融账户持有人的违规行为，通过利用系统的漏洞来达到经济上的收益。

对于支付场景中的风险，第三方支付平台可以通过对一周或者一个月内的账户资金流入流出进行分析，如果资金的流动密集集中在一些账户，而这些账户活跃的 IP、设备是同一个或者相近的，那么风险异常的概率是非常高的，可能存在着盗卡支付和洗钱套现的行为。

6.2 大数据在网络借贷中的应用

6.2.1 推荐系统简述

推荐系统(Recommender System)，是指建立在海量数据挖掘基础上的一种高级商务智能系统，它是一种把用户提供的推荐信息作为输入，然后将这些信息进行聚合、处理，最后把相关信息投放给合适的客户的信息服务。形象地讲，推荐系统就是用来在用户的兴趣与被推荐物品之间所搭起来的一座桥梁。例如，当客户在逛淘宝时，发现淘宝的主界面会出现用户购买过、收藏过、浏览过的商品或相关商品。这种对用户的商品推荐就是由推荐系统运作形成的。

推荐系统主要包括 3 个部分：输入模块、推荐引擎模块、输出模块。

1. 输入模块

输入模块又被称为用户模块，这里的用户指的是信贷产品的购买者和潜在购买者。在系统中，输入模块的主要作用是负责收集和更新用户的信息。具体而言又包括两大部分，一部分是用户的基本特征信息，包括年龄、性别、职业、收入等；另一部分是用户的行为信息，包括显性信息(如评分、评论等)和隐性信息(用户浏览网页的停留时间、点击率、客户转化率等)。

2. 推荐引擎模块

推荐引擎模块又称客户推荐算法模块，它是整个推荐系统的核心部分。该模块通过算法对输入模块所采集的数据进行分析处理，之后再将结果输出。推荐算法的好坏直接影响到整个系统的效率和效果。常见的推荐算法主要有协同过滤推荐、基于内容的推荐、基于知识的推荐等。

(1) 协同过滤推荐(Collaborative Filtering Recommendation)。是推荐系统中应用最早和最为成功的技术之一。它一般采用最近邻技术，利用用户的历史喜好信息计算用户之间的距离，然后利用目标用户的最近邻居用户对商品评价的加权评价值来预测目标用户对特定商品的喜好程度，系统从而根据这一喜好程度来对目标用户进行推荐。协同过滤的最大优点是对推荐对象没有特殊的要求，能处理非结构化的复杂对象，如音乐、电影。

(2) 基于内容的推荐(Content-based Recommendation)。是信息过滤技术的延续与发展，它是建立在项目的内容信息上做出推荐的，而不需要依据用户对项目的评价意见，更多地需要用机器学习的方法从关于内容的特征描述的事例中得到用户的兴趣资料。在基于内容的推荐系统中，项目或对象通过相关的特征的属性来定义，系统基于用户评价对象的特征，学习用户的兴趣，考察用户资料与待预测项目的相匹配程度。用户的资料模型取决于所用的学习方法，常用的有决策树、神经网络和基于向量的表示方法等。基于内容的用户资料需要有用户的历史数据，用户资料模型可能随着用户的偏好改变而发生变化。

(3) 基于知识的推荐(Knowledge-based Recommendation)。在某种程度是可以看作一种推理技术，它不是建立在用户需要和偏好基础上推荐的。基于知识的方法因它们所用的功能知识不同而有明显区别。效用知识是一种关于一个项目如何满足某一特定用户的知识，因此能解释需要和推荐的关系，所以用户资料可以是任何能支持推理的知识结构，它可以是用户已经规范化的查询，也可以是一个更详细的用户需要的表示。

3. 输出模块

输出模块是一个将推荐结果展现给客户的一个模块，它包含多种形式，例如建议、预测、个体以及评分等等。例如互联网用户在“淘宝”主页上发现自己喜欢的商品以及浏览过的商品，这就是一种界面化的建议。

6.2.2 P2P 网站中的个性化推荐

个性化推荐是指根据用户的兴趣特点和购买行为，向用户推荐其可能感兴趣的信息和商品。在 P2P 网站中，随着信贷规模的不断扩大，信贷产品个数和种类快速增长，贷款者往往需要耗费大量的时间和精力才能找到合适的信贷产品。而浏览大量无关的信息和产品无疑会降低用户的使用体验，使淹没在信息过载问题中的消费者不断流失。

为了解决 P2P 网站中的这一问题，个性化推荐系统是一个可行的解决思路。在海量数据挖掘基础上，构建一个基于个性化推荐系统的高级商务智能平台，可以通过电子商务网站为其客户购物提供个性化的决策支持和信息服务。一般 P2P 网站有大量的信贷产品，用户常常感到难以入手，如果有一种信贷产品选购的助手，能根据客户的兴趣爱好推荐其可能感兴趣的信贷产品，可以有效提高客户的满意度。

在 P2P 网站中使用个性化推荐的最大的优点在于：一方面可以获取信贷产品的特点，如贷款人的特征和借款记录，信贷产品的期限、风险和收益度；另一方面又可以获取用户的特点，如客户的产品浏览记录、个性化需求和兴趣偏好、客户的个人属性、客户过去的贷款行为和贷款记录等，从而为贷款客户做出个性化推荐。此外，系统给出的推荐是可以实时更新的，即当系统中的信贷产品库或用户特征库发生改变时，给出的推荐序列会自动改变，这就大大提高了 P2P 贷款的简便性和有效性，同时也提高了 P2P 平台的服务水平。

总体来说，一个成功的 P2P 网站个性化推荐系统的作用主要体现在以下 3 个方面。

(1) 将 P2P 平台的浏览者转化为使用者。一般地，P2P 平台的浏览者在浏览过程中并没有投资或贷款的欲望，而个性化推荐系统能够推荐他们感兴趣的信贷产品，从而形成有效的客户转化。

(2) 提高 P2P 平台的交叉销售能力。个性化推荐系统在贷款客户的选择过程中向用户推荐其他有价值的信贷产品，用户能够从系统提供的推荐列表中找到自己确实需要但在购买过程中没有想到的信贷产品，从而有效提高 P2P 平台的交叉销售。

(3) 提高客户对 P2P 平台的忠诚度。与传统的贷款模式相比，P2P 平台使得用户拥有越来越多的选择，用户更换信贷产品极其方便，只需要点击一两次鼠标就可以在不同的 P2P 平台之间跳转。个性化推荐系统可以分析客户的贷款习惯，根据客户需求向用户提供有价值的信贷产品。高质量的推荐系统可以使用户产生依赖。因此，个性化推荐系统不仅能够为用户提供个性化的推荐服务，而且能促进 P2P 平台与用户建立长期稳定的关系，提高客户忠诚度，防止客户流失。

个性化推荐系统具有良好的发展和应用前景。目前，许多 P2P 网站都不同程度地使用了各种形式的推荐系统。在大数据环境下，个性化推荐系统能够有效地保留客户，提高 P2P 网站的服务能力，为其带来巨大的经济效益。

6.2.3 基于 VITA 系统的信贷产品匹配机制

VITA 金融服务推荐系统是为匈牙利 Fundamental 信贷协会开发的一种基于知识的推荐技术。VITA 能够帮助销售代表与客户在销售过程中进行交互，可以提高销售代表的工作业绩，降低开发和维护相关软件的整体费用，通过该工具，可以构建一个基于知识的推荐技术。其知识获取平台可以通过图形用户界面开发推荐系统知识库和推荐过程定义功能。

这种知识获取平台也适用于 P2P 平台。目前 P2P 遇到的最大挑战就是金融行业信贷产品的增多与难以满足的客户的个性化借贷需求。一方面，客户面临如此众多的信贷产品无所适从；另一方面，P2P 的工作人员也很难为每一个客户选择他们最适合的信贷产品，也很难解释将这些产品推荐给用户的原因。因此，推荐系统的供应商的主要目标是提高推荐的整体工作效率。这就需要提高算法的准确度，同时还要提高客户的黏度，让客户有兴趣使用某个 P2P 平台，这就需要工作人员能够处理极为复杂且频繁变化的推荐知识库。基于知识的推荐技术能够改善这种情况，因为它能高效地挖掘并维护知识库。

将 VITA 系统用于 P2P 的产品推荐，其目标主要有以下两点。

(1) 提高贷款成交数。在相同的时间内，提高成功融资产品的数量。

(2) 有效的软件开发和维护。新技术应该能改善配置知识库的开发工作。

而最终的 VITA 支持平台可自动实现对大量历史交易数据的学习，并实时更新配置知识库，并将新知识应用于信贷产品和客户投资需求的映射。更新知识库时，知识获取平台可以自动测试并调试知识库。该知识库包括以下元素。

(1) 用户属性。每个用户必须表明自己的需求，这是合理推荐的前提条件。在 P2P 服务领域，用户属性的例子有年龄、风险承受能力、预期的放贷有效期、现有贷款组合等。

(2) 产品属性及实例。每个信贷产品都用事先定义好的一组属性词描述，比如，不应该向没有准备好承担风险的用户推荐高风险的产品。

(3) 约束。确定某些场景下不应该向客户推荐某些信贷产品。例如，不应该向风险承受能力差得用户推荐风险高的产品。

(4) 咨询过程定义。对信贷产品推荐规则进行明确的定义，构建状态图(见图 6.1)并以此确定客户提出问题的场景。图中方框内容表示推荐规则的流程，圆圈内容表示客户在每个阶段可能提出的问题。

正如图 6.1 所示，信贷推荐过程包括 4 个阶段：提取需求、信用价值审核、产品咨询和选择、精确计算及展现结果。在第一阶段，系统提取客户的基本信息、信贷的目的和需求。提取之后再审核客户的信用价值，根据当前客户的金融状况、历史借贷记录、金融有价证券、财务状况等详细信息进行评估。这时，系统应检查能否找到满足当前需求的解决方案。如果没有，那么系统户设法找到其他符合客户需求的方案。在信用价值审核之后，推荐系统可能会推荐多种符合客户要求的信贷产品。当客户选择其中一个产品时，推荐系统会计算并提供详细的信贷产品的属性，包括月度偿还率、偿还期限、保留条款等。

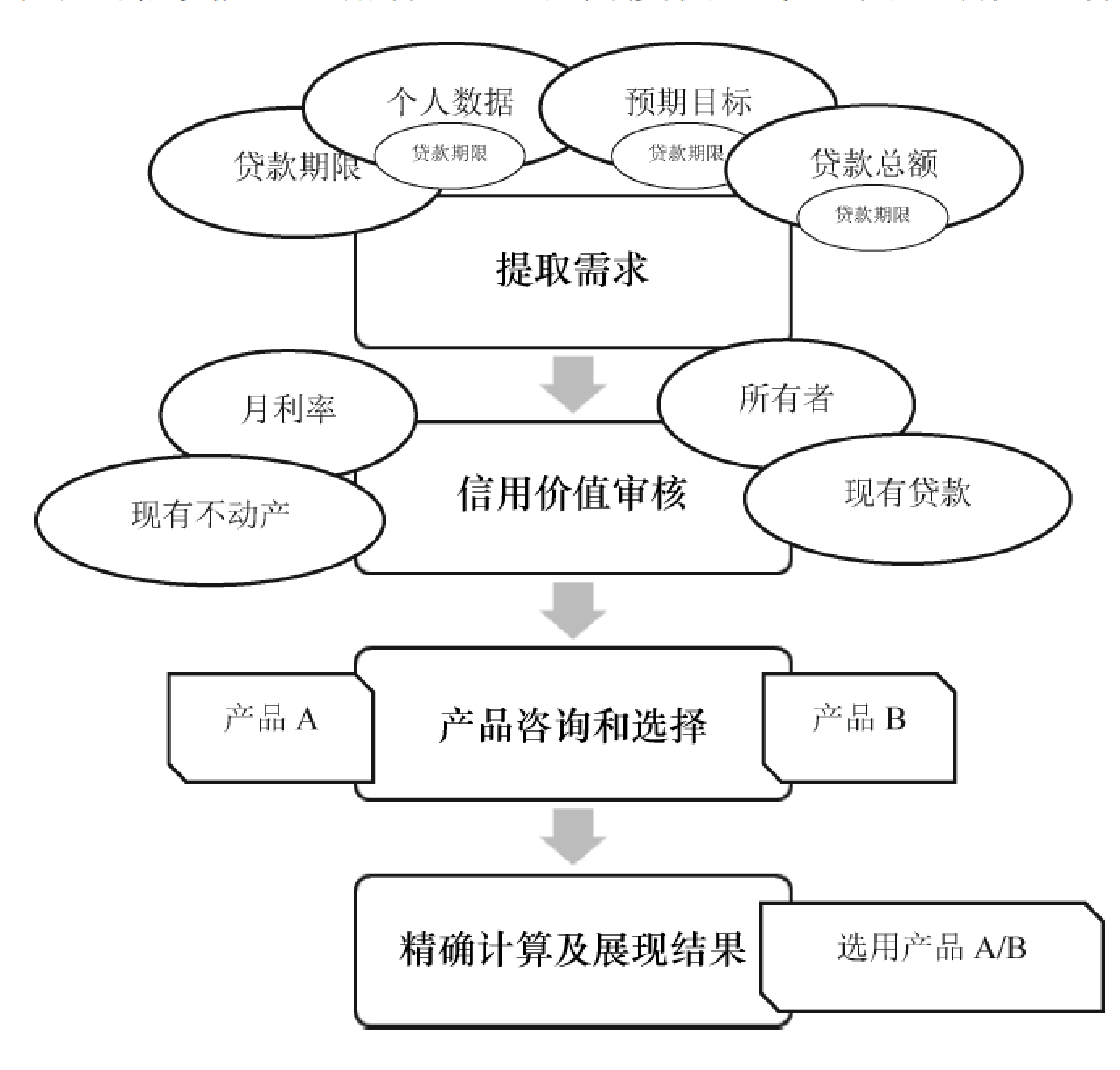


图 6.1 咨询过程定义示例

完成信贷产品与客户的匹配，客户也选择了该信贷产品之后，P2P 平台就可以开始向客户放贷了。在放贷过程中，用户随时可以对信贷产品进行评价，包括还款准时率、收益

率、客户满意度等。推荐系统可以收集这些信息，通过机器学习的方法获取其中的知识，然后自动更新已有的知识库，以准确地实现产品推荐。

6.3 大数据在互联网供应链金融中的应用

在供应链金融服务中，供应链融资(Supply Chain Financing)是最为核心的业务。究其定义，供应链金融是指把供应链上的核心企业及其相关的上下游配套企业作为一个整体，根据供应链中企业的交易关系和行业特点制定基于货权及现金流控制的整体金融解决方案的一种融资模式。供应链融资是对银行传统信贷模式的全面改革，是对贸易融资、企业理财、现金管理等一系列金融产品的整合。

根据资金来源的不同，供应链融资可以分成内部融资和外部融资两种。①供应链内部融资，是指利用上下游企业提供的商业信用来加速资金周转，提高资金利用率；②供应链外部融资，是指利用银行等金融中介机构提供的流动资金贷款来缓解资金约束。然而，在供应链中资金充裕的核心企业，出于自身利益最大化和保持现金流稳定的角度考虑，并不主动愿意为中小企业提供融资，因此需要外部金融机构的介入。

供应链融资解决了上下游企业融资难、担保难的问题，而且通过打通上下游融资瓶颈，还可以降低供应链条融资成本，提高核心企业及配套企业的竞争。

互联网供应链融资(Supply Chain Financing-online)是指利用互联网技术，对供应链中的核心企业提供融资，并通过大数据、云计算控制风险的金融业务。它是一种集成的概念，兼有互联网金融和供应链融资的一般性质，例如互联网金融的便利性、虚拟性、创新性以及供应链融资的流程控制、成本控制、严格的风险管理等。

根据经营主体的不同，互联网供应链融资大致可以分为 3 类。

(1) 合作模式的互联网供应链融资。商业银行与互联网电商通过合作协议的形式，利用商业银行的资金优势及互联网电商的电子商务诚信体系，融合资金流、信息流和物流，向在 B2B 和 B2C 电子商务平台从事交易行为的小微企业提供信用贷款等信贷类产品及现金管理、支付结算等金融服务。例如，中国工商银行(2007 年)、中国建设银行(2008 年)先后与阿里巴巴合作推出的电子商务平台小微企业无抵押贷款即属于此种类型。这是最早产生的互联网供应链融资模式，亦是带有过渡性质的模式。随着下面两类模式的出现，该类模式逐渐消失，因而不能代表互联网供应链融资的未来发展方向。

(2) 电商主导的互联网供应链融资。互联网电商利用其自身的注册资本及电子商务诚信体系向其 B2B 和 B2C 电子商务平台的小微企业提供信用贷款。例如，2010 年 6 月成立的“阿里小贷公司”即是首家全国范围内的小额贷款公司。

(3) 商业银行主导的互联网供应链融资。商业银行自主建立 B2B 和 B2C 电子商务平台，同时兼具了电商和资金提供者的身份。一方面，为中小企业或小微企业提供交易信息发布、在线交易的电子商务平台渠道；另一方面，在全方位掌握在线企业交易信用数据的基础上，建立电子商务诚信体系，向企业提供支付结算、融资贷款、资金托管等全方位的专业服务。例如，建设银行的“善融商务企业/个人商城”(2012 年)和工商银行的“融 e 购企业商城/个人”(2014 年)均是商业银行自主建立的 B2B 和 B2C 电子商务平台，直接为平

台上的企业提供供应链融资服务。

6.3.1 基于大数据的互联网企业信用评估

1. 供应链中的企业信用问题

中小型企业对我国的经济发展起着重要的作用，它创造了将近 60%的 GDP，解决了全国 80%的就业。但由于我国的金融市场的不完善，中小企业的融资渠道和手段有限，主要还是靠自有资金和信贷资金维持生产经营。而仅靠自有资金难以支撑企业的发展，因此信贷资金成为中小企业扩大再生产的一种手段。但是由于企业估值不合理、财务报表披露不到位、信用观念薄弱等因素，银行出于风险考虑并不愿意向中小企业提供贷款，因此，融资难一直是中小企业发展过程中的难题。

而供应链融资是缓解中小企业融资难的有效途径。在供应链融资中，正确评估企业的资信是控制供应链融资风险的核心内容，即提高金融机构对中小企业信用风险评估的准确性，使优良的中小企业及时得到贷款，同时降低金融机构所面临的信用风险。

2. 基于人工智能的信用评分模型

金融机构为了降低互联网供应链金融业务中的信用风险，常会借助统计学的方法确定借款者的信用度，并通过科学的算法建立信用评分模型，从而将企业的信用状况量化成为信用评分。

信用评分模型已经被金融机构普遍采用。金融机构通过信用评分模型降低贷款程序中的开销，减少不良贷款带来的损失，从而为有效的决策提供强有力的支持。

信用评分的发展大致经历了专家分析、统计分析和人工智能 3 个阶段。1970 年以后，信用评分主要采用定性的方法，主要的分析方法有 5C 分析法、五级分类分析法。大多数金融机构基本上是靠专家的分析来评估信用风险。但是，随着业务量的增大，人工完成这项工作是不可能的。因此，信用产业的许多机构纷纷开发新的模型来支持信用决策。这些模型旨在提高衡量信用程度的准确率，从而降低信用风险、减少损失。1970—1990 年，金融机构主要采用的是基于财务指标的信用评分模型。第一个信用评分模型由 Altman 提出。现代的信用评估模型可以分为两种：统计模型和人工智能模型。最常见的统计模型为线性判别分析(LDA)和 Logistic 回归(LR)。由于变量之间的线性关系不足，因此这两个模型的准确度不高。而随着信息技术的发展，一些人工智能的方法已经被用来建立准确且稳定的信用风险评估系统。例如，人工神经网络(ANNs)、决策树(DT)、贝叶斯分类器(BC)、模糊规则系统以及集成学习模型等，在信用风险评估中取得了良好的效果。与统计模型不同的是，人工智能可以直接从数据中集中获取训练知识，并不需要关于变量分布的假设。因此，人工智能模型的性能更好。

3. 基于 PSO-BP 集成的企业信用分

关于大数据对企业信用的评估，下面简单介绍较为典型的 PSO-BP 集成的企业信用评分模型。该模型的流程有以下几个步骤。

第一步，使用 bagging 抽样技术获得足够多不同的训练数据集。

第二步，构建 PSO-BP 组合成员分类器，然后使用不同的训练数据集训练此分类器。

第三步，使用组合投票准则整合不同组合成员分类器的分类结果，得到企业的信用分。

第四步，在测试数据集上测试模型的性能。

该模型的整体架构如图 6.2 所示。

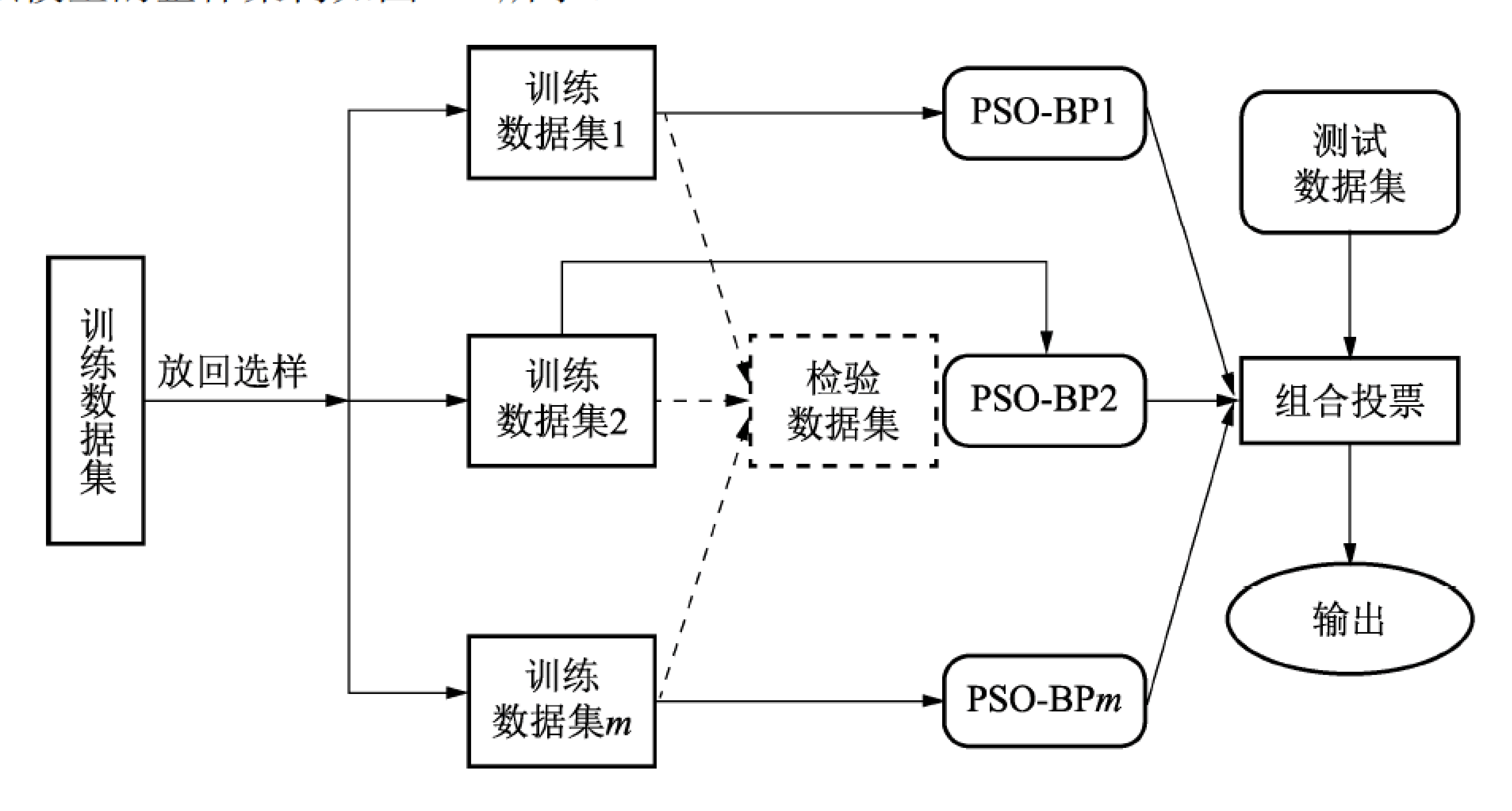


图 6.2　PSO-BP 集成模型整体架构

1)　产生训练集子集

现实中，当需要对一个重要问题做决策时，往往需要综合多个专家的意见。在机器学习中也一样，常常需要构建多个计算模型，然后综合所有模型的运算结果得出结论。Bagging 算法是机器学习领域中广泛使用的数据抽样算法，常被用来从原始数据集中创建不同的样本，用以得到不同的分类计算模型。它作为机器学习领域极为有效的数据处理模型，采用随机放回抽样的方式，可以得到一定数量的训练数据集。

2)　创建不同的分类器

组合模型能取得更高分类准确率的一个充分必要条件为，用于组合的分类器必须是准确和有差异的。一般地，组合成员差异度较大的组合模型具有更高的泛化能力。因此，如何生成差异度最大的分类器成为一个关键的问题。对神经网络模型来说，为生成不同的分类器模型，可以通过参数变化、网络结构设计的变化或是神经网络训练方式的变化等方法实现。

3)　训练 PSO-BP 模型

由于 PSO-BP 模型具有泛化能力和收敛速度上的优势，所以选用该模型。PSO-BP 模型是 PSO 与 BP 神经网络的混合优化算法。该混合算法的根本出发点在于，在初始阶段使用 APSO 进行全局搜索，然后使用 BP 在全局最优位置附近进行局部搜索，从而提高收敛速度。在 PSO-BP 的使用过程中，考虑到 PSO 的迭代次数较少(5 次)，采用惯性权重随着算法迭代自动变化的 APSO 算法意义不大，因此选择了带压缩的粒子群算法进行迭代寻优。PSO-BP 算法的流程如图 6.3 所示。

(1)　在[0，1]范围内随机初始化粒子群体的位置和速度。

(2) 计算每个粒子的适应值，初始为当前粒子的局部位置，设置为初始种群的全局最优位置。

(3) 如果进入最大的迭代次数，算法转到(7)，否则继续运行(4)。

(4) 存储当前种群的最优粒子，并更新粒子的速度和位置，这样就形成了一组新的种群，如果新的粒子位置超出了界限[X_{min}， X_{max}]，新的位置将会被设置为 X_{min} 或 X_{max}；如果新的粒子速度超出了界限[V_{min}, V_{max}]，新的速度将会设置为 V_{min} 或 V_{max}。

(5) 计算每个粒子的适应值，最差的粒子由最好的粒子代替。如果粒子 i 的新位置比 P_{ib} 更好，该位置将作为第 i 个粒子新的 P_{ib}。如果所有粒子中存在比 P_g 更优的粒子，P_g 将得到更新。

(6) 使用 BP 算法在 P_g 附近搜索，如果搜索结果由于 P_g，将 P_g 作为当前搜索结果。另外，在搜索过程中，如果测试数据集上 P_g 超过一定的迭代次数没有变化，BP 算法搜索结束。

(7) 输出全局最优的 P_g。

4) 分类结果集成

基于上述几个步骤的工作。可以得到一组不同的 PSO-BP 组合分类成员。之后就是通过一个适当的组合策略将不同的分类器集成为一个分类器，常见的组合策略包括多数投票准则、排序准则和权值平均等。最常用的就是多数投票准则，在该策略中，组合成员分类器的成员决定了最终的输出。

总之，基于 PSO-BP 集成的信用评估模型(见图 6.3)，与其他的信用评估模型相比准确率更高。

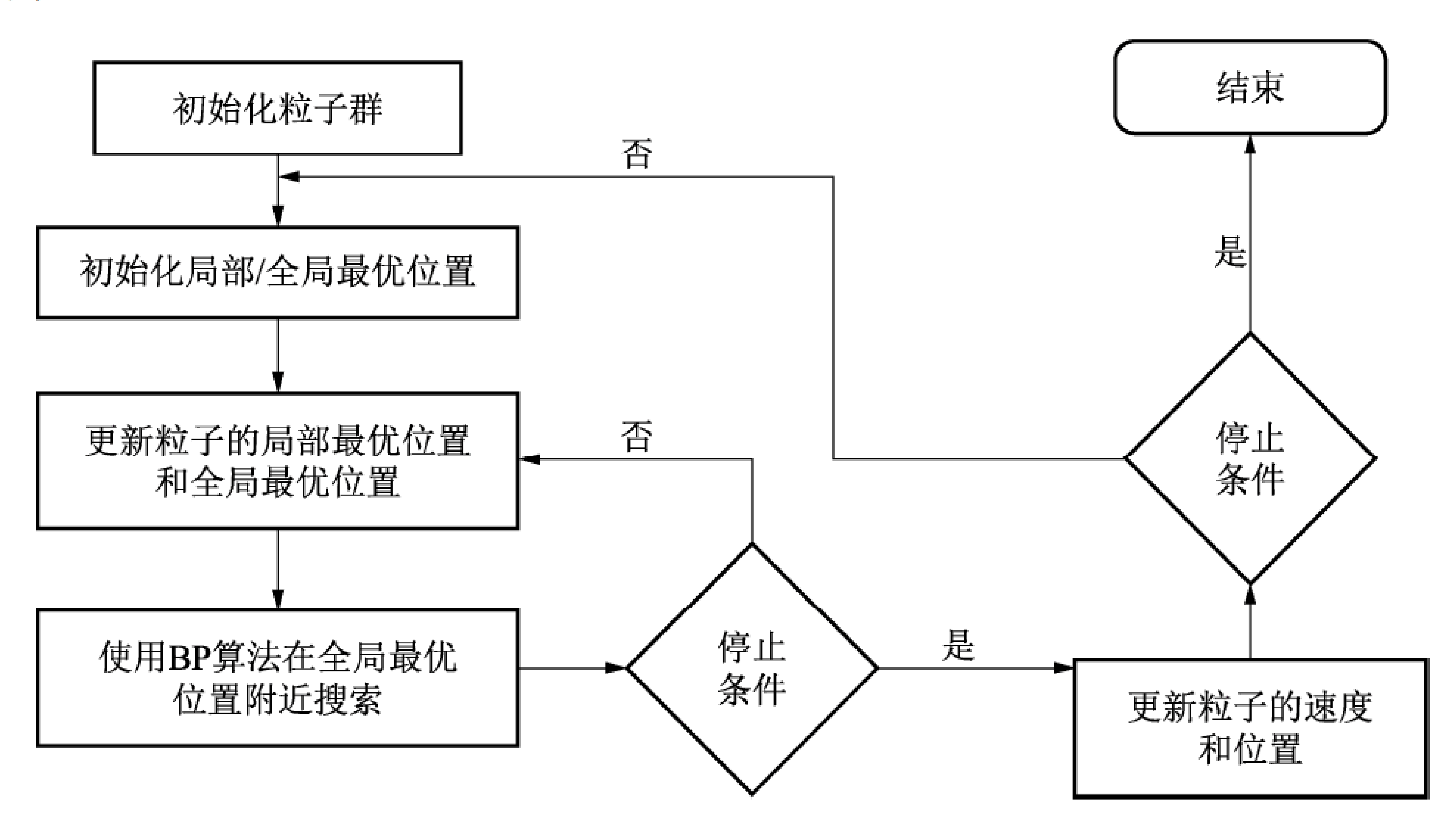


图 6.3 PSO-BP 算法的流程

6.3.2 案例：京东供应链金融模式

近年来，京东频频加码互联网金融，供应链金融是其金融业务的根基。京东通过差异化定位及自建物流体系等战略，并通过多年积累和沉淀，已形成一套以大数据驱动的京东供应链体系，为上游供应商提供贷款和理财服务，为下游消费者提供赊销和分期付款服务。具体可以分为采购订单融资、入库环节入库单融资、结算前应收账款融资、担保、保单业务扩大融资、协同投资信托计划、资产包转移计划、消费者分期付款、消费者投资理财等类型，涉及应收账款融资、订单融资、委托融资、协同融资、信托计划、京东白条、校园白条、保险、理财、黄金等产品。京东有非常优质的上游的供应商、下游的个人消费者、精准的大数据，京东的供应链金融业务水到渠成。京东商城的 CEO 刘强东也表示，未来的商业竞争是供应链的竞争，而供应链金融提高了供应链整体的运营能力，通过资金流带动整个链条不断向前滚动，从而实现供应链的有机整合。京东供应链金融模式的具体流程如图 6.4 所示。

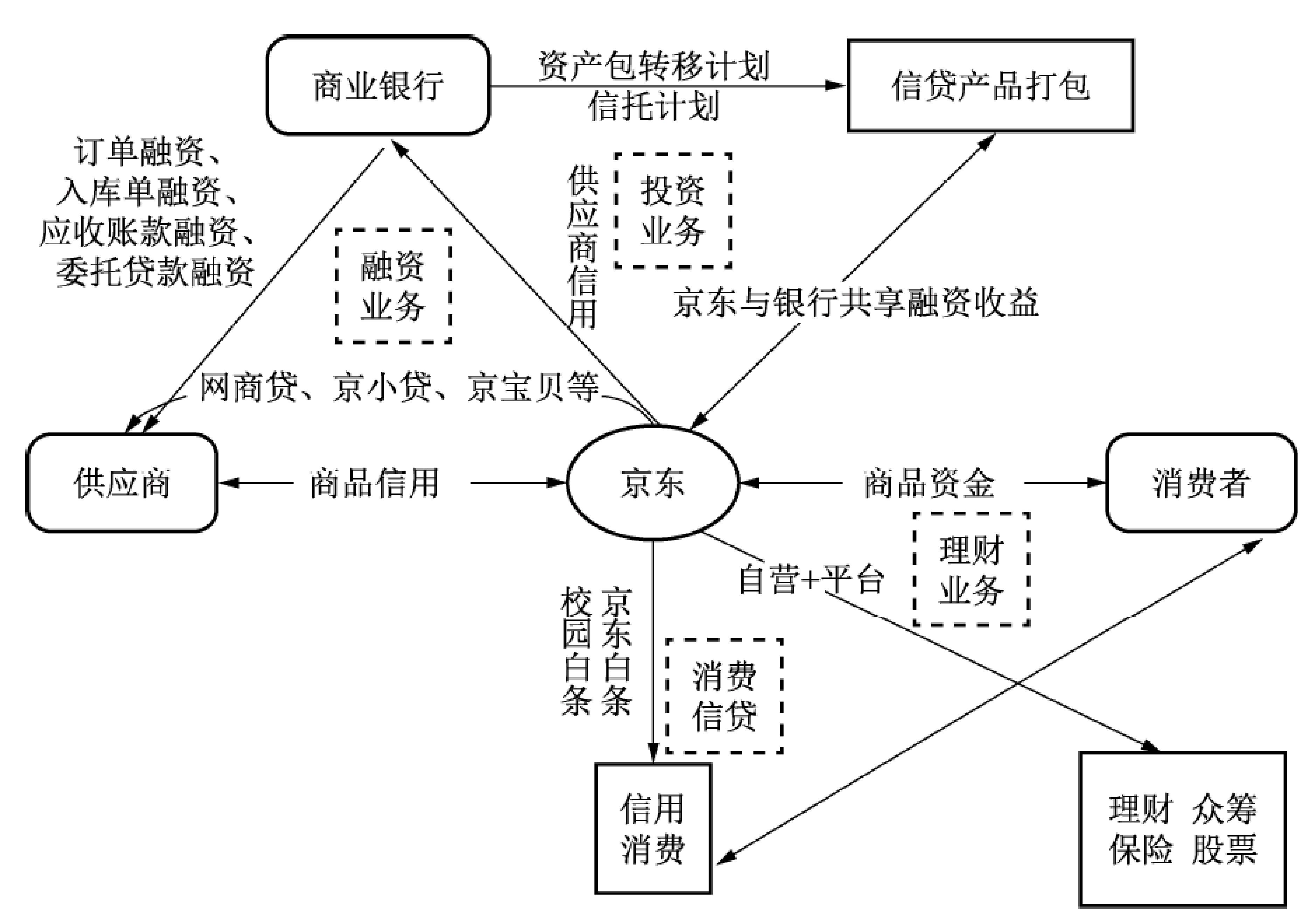


图 6.4 京东供应链金融模式的流程

京东供应链金融有个很有时代意义的创新产品——动产融资。传统动产融资有三大困局：一是抵质押物范围小，广大中小企业的动产价值难以评估并用来质押；二是缺乏全国性的、电子化的动产质押登记平台，导致重复质押等风险事件频发；三是质押方式死板，货物一旦用来融资，流动性将大大降低，不能随着买进卖出自动调整融资额度。以上问题导致大量中小企业不能被纳入动产融资服务的覆盖范围。京东供应链金融设计了一款可以同时解决以上问题的底层架构，可以通过数据和模型化的方式自动评估商品价值，他们与具有“互联网+”特点的仓配企业合作，全面整合了质押商品从生产、运输、存储到销售的全链条数据交叉验证，实现动态质押。这款产品一经推出，就迅速实现单月放贷破亿。

动产融资还在向 B2B 平台采购的经销商提供服务。

基本上，京东供应链金融的创新产品已经在为各类场景、特点的企业服务，覆盖了很多传统融资触达不到的群体。2015 年年底，京东供应链金融宣布将与企业理财打通，让企业缺钱时借钱，有钱时理财。作为一家仅仅提供供应链金融 3 年多的企业，京东供应链金融的创新速度和展业速度令人吃惊。

6.4 大数据在互联网消费金融中的应用

6.4.1 互联网消费金融的大数据征信与风控

消费金融中最重要的一个问题就是本身所存在的信用风险。由于消费金融的主要客户群体是年轻人群和中低收入人群，且主要是以个人的信用状况为担保，没有抵押物，因此，经营主体出于风险管理角度的考虑，对客户的征信非常重要。如果按传统的方法依靠线下收集客户信息来判断其还款能力和还款意愿，不光效率低下无法获得尽可能多的客户，还无法对客户进行有效的信用评估。

另外，就目前我国实际的情况而言，大多数消费金融经营机构没有丰富的征信经验和征信能力，大数据征信产业由此产生。许多消费金融公司依托大数据征信机构。这些机构利用大数据的方式收集客户信息，通过对客户群体的消费数据分析，进行客户评级，获得有效的风控模型，进而对客户进行分流和筛查，进行差异化管理，并不断优化风控模型和信贷审核流程，达到可量化的自动化决策的目的。

面对蓬勃发展的互联网消费金融的风控需求，恒生电子推出了大数据风控平台，为中小型消费金融厂商提供强大的风控服务，从以下 3 个方面提供专业的大数据风控支持。

(1) 外部数据源整合。整合第三方数据源与征信服务机构，从反欺诈、证据保全到第三方征信、电商平台等多维度全方位的数据与服务。

(2) 风控模型与评分。从还款能力与还款意愿等多角度对客户进行审核，对不同种类的客户进行差异化评估，并基于评分卡进行审批、授信、差异化定价、风险预警、额度调整等流程的设计，实现信贷工厂的批量化与规模化的要求。

(3) 自动化决策。针对互联网消费金融的快速放贷的要求，搭建了一套自动化决策模型和风控体系，进行欺诈风险的评估，计算信用风险等级，并给出可信任的参考授信额度，达到快速授信、实时放贷的目的。

那么，如何对互联网消费金融实现有效的大数据风控呢？

首先，运用大数据技术对互联网消费金融业务进行风控，一定要与消费场景相结合，把大数据风控植入一个个消费场景中。消费金融的场景化有助于明确贷款的实际用途，避免了贷款挪为他用所造成的风险。除了网购之外，教育培训、旅行、租房、购车、婚庆、美容等 O2O 场景都具有良好的消费金融属性。不同的场景有不同的用户群，消费金融公司需要设计各不相同的消费金融产品和制定有针对性的贷款政策，而利用大数据技术可以通过数据的采集与分析、各个消费场景和消费群体的特点，确定差异化的贷款政策。

再次，还要加强对网络欺诈的重点防控。互联网消费金融具有互联网的特殊性，一般

为纯线上交易，容易被不法团伙所利用，产生盗号、套现等欺诈行为，且网络欺诈作案手段隐蔽、形式多样，扩散也极快，对风险控制提出了很高的要求。

最后，可以运用多种风险分散手段，如与保险和担保机构合作。保险机构在提供各种信用保证保险产品的同时，也可将其自身的征信服务提供出来；担保公司由于在风险防范机制上比较专业，可用来完善消费金融公司自身的风控模型。

6.4.2 案例：芝麻信用

芝麻信用，是蚂蚁金服旗下独立的第三方征信机构，通过云计算、机器学习等技术客观呈现个人的信用状况，已经在信用卡、消费金融、融资租赁等上百个场景为用户、商户提供信用服务。它运用大数据技术，从个人用户的信用历史、行为偏好、履约能力、身份特质和人脉关系 5 个维度，对个人信用予以评价并将其量化成为芝麻信用分。芝麻信用分越高，代表用户的信用状况越好。

(1) 在数据来源方面。芝麻信用除了使用强大的淘宝、天猫电商数据以及支付宝金融数据之外，还涵盖了信用卡还款、网购、转账、理财、水电煤缴费、租房信息、住址搬迁历史、社交关系等。用户信用分的高低与网购量、财产多少没有直接联系，而是与他平时的守信程度有关。此外，芝麻信用还与公安网等众多公共机构有深入的数据合作关系，同时也将开辟各类渠道允许用户主动提交各类信用相关信息。

(2) 在数据算法方面。芝麻信用体系将包括芝麻分、芝麻认证、风险名单库、芝麻信用报告、芝麻评级等一系列信用产品，背后则是依托阿里云的技术力量，对 3 亿多实名个人、3700 多万户中小微企业数据进行整合。借助阿里云，不论是从算法准确率上来说还是从安全、稳定等多个方面来讲，芝麻信用都具有非常优越的领先条件。

(3) 在获取用户入口方面。阿里芝麻信用通过依托于支付宝平台，很快就获得了快速稳定的用户增长，这个优势是其他平台所不具备的。依托于支付宝，凌驾于淘宝、天猫等购物商城之上，芝麻信用很快就推出了蚂蚁花呗、蚂蚁借呗等信用产品，并与招联金融旗下的“好期贷”达成了战略合作，全面进军消费金融领域。蚂蚁借呗 3 秒钟便可完成放贷，最高可以获得 5 万元的消费贷款，钱直接从余额宝余额转出，用途不限，非常方便。

本章总结

- 大数据技术在第三方支付行业中主要是应用于防范欺诈风险，第三方支付公司通过 IT 技术建立云端的动态数据库，收纳客户的基本信息，并通过一定的算法建立风险控制模型。在第三方支付的注册场景、登录场景、绑卡场景以及支付场景中大数据技术在欺诈风险控制方面发挥着重要的作用。
- 网络借贷平台主要利用大数据技术中的一些重要的算法构建推荐系统， 例如协调过滤推荐、基于内容的推荐以及基于知识的推荐等。许多 P2P 网站都不同程度地使用各种形式的推荐系统，从而提高网站的服务能力，有效地发展及保留了客户。

- 在供应链金融中，大数据技术主要应用于企业的信用评估，通过构建信用评分模型来对企业的信用状况进行评价，从而确定企业合理的授信额度。目前一些人工智能的方法已经被应用到该领域，例如人工神经网络、决策树、贝叶斯分类器等。PSO-BP 模式是最为典型的一个信用评估模型，与其他的信用评估模型相比它的准确率较高。
- 在互联网消费金融行业中，大数据技术主要应用于征信以及风险管控。由于消费金融具有无抵押、以信用为基础的特点，因此信用风险是一个非常重要的问题。通过外部数据源整合、风控模型与评分、自动化决策等大数据技术为互联网消费金融机构提供强有力的风控支持。

本章作业

1. 第三方支付的欺诈风险主要体现在哪些方面？你认为第三方支付机构在欺诈风险管理方面应该承担哪些责任？
2. 第三方支付中注册场景、登录场景、绑卡场景以及支付场景中的欺诈风险的具体表现形式是什么？如何利用大数据技术在这 4 个场景中防范欺诈风险？
3. 什么是推荐系统？它主要包括哪些模块？举例说明日常经济生活中的推荐系统。
4. 试述推荐系统中几个重要的推荐算法。
5. 个性化推荐系统对 P2P 平台有哪些作用？
6. 什么是供应链融资？什么是互联网供应链融资？根据经营主体的不同，互联网供应链融资可以分为哪几类？
7. 简述基于 PSO-BP 集成的企业信用分的产生过程。
8. 阅读本章 6.3.2 小节的案例，简述京东是如何将大数据技术运用到互联网供应链融资中的？
9. 如何对互联网消费金融进行大数据征信与风控？
10. 阅读本章 6.4.2 小节的案例，试述芝麻信用是如何利用大数据进行征信和风控的？

第7章

大数据征信

本章目标

- 理解并掌握传统征信的含义、原则、分类及特征；了解传统征信的基本流程
- 掌握传统征信产品、机构及体系
- 理解并掌握大数据征信的含义、优势及难题；了解其理论基础，并知晓大数据征信出现的必然性
- 掌握大数据征信流程；了解国内外典型大数据征信企业

本章简介

数据是征信业务开展的基础资料。征信活动主要是围绕数据进行采集、整理、保存、加工，并最终向信息使用者提供。大数据不仅为征信业发展提供了极为丰富的数据信息资源，也改变了征信产品设计和生产理念，成为未来征信业发展最重要的基石。

我国征信业发展尚处于起步阶段，在大数据时代存在征信法律制度和业务规则不够完善、征信机构数据处理能力有待提高等问题。未来征信业面临的机遇和挑战并存，研究大数据时代征信业的发展具有重要意义。

本章从传统征信入手，详细阐述了传统征信的含义、原则、分类及特征，大致介绍传统征信的基本流程，并全面叙述传统征信产品、机构及体系；进而讲述了大数据征信的含义，并与传统征信对比阐述其优势和难题，从各个角度说明大数据出现的必然性；最后以典型大数据征信企业作为突破口，立体叙述了在实践中的大数据征信流程，并对典型大数据征信企业的运作和征信模式做了大致的介绍。

7.1 传统征信

7.1.1 征信概述

19 世纪初，英国常有“绅士不付裁缝账”的现象，伦敦的裁缝们为绅士和贵族定做衣服是做好之后再收钱，结果总有一些客户不及时付款或故意赖账，这样给裁缝们造成了很大损失。于是，为保护自身利益，裁缝们创立了一个交流其客户支付习惯信息的机制，拒绝为那些信用不良的客户们服务。从这个征信制度的雏形可以看出，征信活动是在授信人之间形成一种分享客户信用信息的机制。

随着市场经济的发展，授信活动或信用活动在市场交易中日益频繁。全社会特别是授信人、投资人对征信服务的需求也不断增长，征信业开始在世界各地蓬勃发展起来。相对于“社会信用体系”“诚信体系”等词来说，“征信”的概念在国际上是有共识的。“征信”对应英语中最合适的词是 credit reference，这里的“信”即“信用”，credit 是指经济层面上的信用。

1. 征信的含义

征信是指征信机构作为信用交易双方之外的独立第三方，收集、整理、保存、加工个人、法人及其他组织的信用信息，以在一定程度上揭示信息主体的信用风险状况，协助授信人或投资人进行风险管理的一种信息服务活动。简而言之，征信的本质就是为授信机构或投资人的决策提供信息参考，是授信人或投资人之间的一种信息分享机制。

这一定义包含了 4 个方面的主要内容。

(1) 这里的信用交易是广义的，是指任何购买(商品和服务等)支付不是同时进行的交易。

(2) 这里的第三方机构就是征信服务机构。而与征信服务相关的服务产品、价格、市场以及征信服务的主体、法规管理等之和，就是征信体系。换句话说，征信体系指的是与征信活动有关的法律规章、组织机构、市场管理等共同构成的一个体系。

(3) 征信体系的主要功能是为借贷市场服务，但也可服务于商品市场和劳动力市场，只要有信用发生，授信方就有征信需求。

(4) 这是一个特殊的信息服务业。它的特殊性主要表现在两个方面：一是信息的特殊性，即它是反映信息主体(企业或个人)信用状况的信息；二是功能的特殊性，除了直接为授信机构提供的服务功能，还具有促进全社会珍惜自己的信用状况、注重诚实守信等延伸的社会功能，有利于构建和谐社会。

2. 征信的原则

征信的原则是征信业在长期发展过程中逐渐形成的科学的指导原则，是征信活动顺利开展的根本。通常，我们将其归纳为真实性原则、全面性原则、及时性原则及隐私和商业秘密保护原则。

1) 真实性原则

真实性原则，是指在征信过程中征信机构应采取适当的方法核实原始资料的真实性，以保证所采集的信用信息是真实的，这是征信工作最重要的条件。

只有信息准确无误，才能正确反映被征信人的信用状况，保证对被征信人的公平。真实性原则有效地反映了征信活动的科学性。征信机构应基于第三方立场提供被征信人的历史信用记录，对信用报告的内容，不妄下结论，在信用报告中要摒弃含有虚伪偏袒的成分，以保持客观中立的立场。基于此原则，征信机构应给予被征信人一定的知情权和申诉权，以便能够及时纠正错误的信用信息，确保信用信息的准确性。

2) 全面性原则

全面性原则又称完整性原则，是指征信工作要做到资料全面、内容明晰。

被征信人，不论企业或个人，均处在一个开放性的经济环境中。人格、财务、资产、生产、管理、行销、人事和经济环境等要素虽然性质互异，但都具有密切的关联，直接或间接地在不同程度上影响着被征信人的信用水平。不过，征信机构往往搜集客户历史信用记录等负债信息，通过其在履约中的历史表现，判断该信息主体的信用状况。历史信用记录既包括正面信息，也包括负面信息。正面信息是指客户正常的基础信息、贷款、赊销、支付等信用信息；负面信息是指客户欠款、破产、诉讼等信息。负面信息可以帮助授信人快速甄别客户信用状况，正面信息能够全面反映客户的信用状况。

3) 及时性原则

及时性原则，是指征信机构在采集信息时要尽量实现实时跟踪，能够使用被征信人最新的信用记录，反映其最新的信用状况，避免因不能及时掌握被征信人的信用变动而为授信机构带来损失。

信息及时性关系到征信机构的生命力，从征信机构发展历史看，许多征信机构由于不能及时更新信息，授信机构难以据此及时判断被征信人的信用风险，而导致最终难以经营下去。目前，我国许多征信机构也因此处于经营困境。

4) 隐私和商业秘密保护原则

对被征信人隐私或商业秘密进行保护是征信机构最基本的职业道德，也是征信立法的主要内容之一。

征信机构应建立严格的业务规章和内控制度，谨慎处理信用信息，保障被征信人的信用信息安全。在征信过程中，征信机构应明确征信信息和个人隐私与企业商业秘密之间的界限，严格遵守隐私和商业秘密保护原则，才能保证征信活动的顺利开展。

3. 征信的分类

征信的分类如图 7.1 所示。

1) 按业务模式可分为企业征信和个人征信两类

企业征信主要是收集企业信用信息、生产企业信用产品的机构；个人征信主要是收集个人信用信息、生产个人信用产品的机构。

有些国家这两种业务类型由一个机构完成，也有的国家是由两个或两个以上机构分别完成，或者在一个国家内既有单独从事个人征信的机构，也有从事个人和企业两种征信业

务类型的机构，一般都不加以限制，由征信机构根据实际情况自主决定。美国的征信机构主要有 3 种业务模式。

(1) 资本市场信用评估机构，其评估对象为股票、债券和大型基建项目。

(2) 商业市场评估机构，也称为企业征信服务公司，其评估对象为各类大中小企业。

(3) 个人消费市场评估机构，其征信对象为消费者个人。

按业务模式分类

企业征信：收集企业信用信息，生产企业信用产品的机构

个人征信：收集个人信用信息、生产个人信用产品的机构

按服务对象分类

信贷征信主要服务对象是金融机构，为信贷决策提供支持

商业征信主要服务对象是批发商或零售商，为赊销决策提供支持

雇用征信主要服务对象是雇主，为雇主用人决策提供支持

其他征信：诸如市场调查，债权处理，动产、不动产鉴定等

按征信范围分类

区域征信一般规模较小，只在某一特定区域内提供征信服务

国内征信是目前世界范围内最多的机构形式之一，尤其是近年来开设征信机构的国家普遍采取这种形式

跨国征信：为拓展业务及顺应国际化趋势，近年来发展迅速

按征信用途分类

公共征信是指出于社会管理需要，征信结果免费提供给社会、政府职能部门、行业协会、商会、联盟开展的征信

非公共征信是指征信用于自己授信和业务管理，其征信过程不公开，自产自销

准公共征信即专业征信，是独立第三方开展的中介服务，其征信结果供社会查询使用，具有社会影响力

图 7.1 征信的分类

2) 按服务对象可分为信贷征信、商业征信、雇用征信及其他征信

信贷征信主要服务对象是金融机构，为信贷决策提供支持；商业征信主要服务对象是批发商或零售商，为赊销决策提供支持；雇用征信主要服务对象是雇主，为雇主用人决策提供支持；另外，还有其他一些征信活动，诸如市场调查，债权处理，动产、不动产鉴定等。

各类不同服务对象的征信业务，有的是由一个机构来完成，有的是在围绕具有数据库征信机构上下游的独立企业内来完成。

3) 按征信范围可分为区域征信、国内征信、跨国征信

区域征信一般规模较小，只在某一特定区域内提供征信服务，这种模式一般在征信业刚起步的国家存在较多，征信业发展到一定阶段后，大都走向兼并或专业细分，真正意义上的区域征信随之逐步消失；国内征信是目前世界范围内最多的机构形式之一，尤其是近年来开设征信机构的国家普遍采取这种形式；跨国征信这几年正在迅速崛起，此类征信之

所以能够得以快速发展，主要有内在和外在两方面原因：内在原因是西方国家一些老牌征信机构为了拓展自己的业务，采用多种形式(如设立子公司、合作、参股、提供技术支持、设立办事处等)向其他国家渗透；外在原因主要是由于世界经济一体化进程的加快，各国经济互相渗透，互相融合，跨国经济实体越来越多，跨国征信业务的需求也越来越多，为了适应这种发展趋势，跨国征信这种机构形式也必然越来越多。但由于每个国家的政治体制、法律体系、文化背景不同，跨国征信的发展也受到一定的制约。

4) 按征信用途可分为公共征信、非公共征信、准公共征信

公共征信是指出于社会管理需要，征信结果免费提供给社会、政府职能部门、行业协会、商会、联盟开展的征信。非公共征信是指征信用于自己授信和业务管理，其征信过程不公开，自产自销，其实质是自我信用风险管理和控制，银行信贷授信、企业信用销售中对客户授信都属于这类。准公共征信即专业征信，是独立第三方开展的中介服务，其征信结果供社会查询使用，具有社会影响力。

4. 征信的特征

1) 征信采集的主要是信用信息

这是征信的第一特征。

信用信息是指能够在一定程度上反映信息主体信用状况的信息。其中，最主要的是与信用交易相关的信息，如贷款、还款信息及合同履约信息等；另一类必不可少的信息，是识别、定位信息主体身份的信息，如名称、身份证件及其代码、地址、年龄、性别，等等。

世界上大的征信机构所建立的征信系统都会采集三类信息：身份识别信息、信贷交易信息和非银行信用信息。作为信用报告主体的信贷交易信息和非银行信用信息主要都是与信用交易相关的信息。其他非银行信息，如法院判决信息、欠税信息、行政处罚信息等，只要是有助于反映信息主体信用状况的，并且法律不禁止，也都是可以采集的。

2) 征信需要建立个人或企业的信息账户

信息账户是征信活动的核心和基础，通俗地说就是一个企业或个人的信用信息档案，即把一个信息主体在各行各业同其他市场主体的信用交易活动中产生的信用记录都整合到一个账户之下。

在我国，最早的企业信用信息档案可追溯到 20 世纪 90 年代，深圳人民银行推出的纸质“贷款证”。人民银行在推广深圳“贷款证”制度并借鉴国外经验的基础上，建立起的全国集中统一的企业和个人征信系统，就是为有信贷交易活动的企业、其他组织和个人建立的信息账户数据库。在信息账户中，信息是需要不断更新的，这是征信系统价值的核心。

3) 征信服务主要是一种微观的信息中介服务

征信具有微观性，在征信活动的两端都表现得很清楚。

从数据采集环节看，征信就是尽可能全面地把信息主体在各行业授信服务、消费和投资活动中留下的信用记录，形成微观经济活动主体——企业和个人的信用报告。从信息使用环节看，通过接受征信服务，商业银行有效地加强了对信贷业务、信用卡和授信客户的

信用风险管理。所以，无论从企业征信还是个人征信的角度来看，征信服务都主要是一种微观的信息中介服务，而不是宏观的信息服务活动。

当然，征信也可以在微观账户数据基础上进行汇总统计和分析，为宏观经济金融分析服务，但这是征信服务的附加产品或增值服务。征信服务应与行政部门的统计服务职能区分开来，不应发生冲突。

4) 征信是一个行业

由征信的发展史可以看出，征信就是从市场经济这个环境中自然而然地孕育出来，为解决交易双方信息不对称问题的专业化服务。也正因为有市场需求，征信才会发展成为一个专门的特殊的信用信息服务行业。

在我国，征信基本上是从银行“贷款三查”(贷前调查、贷中审查和贷后检查)工作中分离出来的，是银行业分工进化的产物。征信机构每天为世界各地的授信机构提供数以百万计的各类征信服务，除信用报告外，还提供包括评分模型开发、防欺诈解决方案、策略决策引擎服务、信息技术解决方案、市场营销服务等。

5) 征信活动的主要服务对象是授信机构

征信的实践活动表明，不仅征信的主要服务对象是授信机构，而且给征信机构提供原始数据的，也主要是授信机构。也正因为授信机构有强烈的需求，才会有动力与征信机构建立起长期、稳定的数据报送关系。

市场经济中授信活动普遍存在，不限于一两个行业。授信机构目前在我国主要是指商业银行，但授信并不是银行的特权。除银行以外，还有很多其他的机构，如小额贷款公司、公积金中心、电信公司等授信机构。理论上，只要是属于先消费或先取货，后付款的交易，都是信用交易或授信活动。在一些欧美国家，授信机构的范围还要广泛，如有些大的超市也可以发信用卡。

6) 征信是一种信息分享机制

这种信息分享机制，只是一定范围内的共享，并不能等同于信息的无限制的、向社会公众的公开披露。就是说，征信产品的使用，即便在授信机构之间，也是有限制的使用，通常是依法依规使用。这是由信用信息的性质决定的，因为它是商务信息，是反映信息主体信用状况的敏感信息。

虽然人们对不同敏感程度的信息应该在什么范围内分享会有不同的认识，但不能由此把范围较大的分享理解为信息公开。尤其是在目前征信系统建设的初期，因缺少相关法律法规的指引，人们对信息分享的范围和参与分享的主体种类存在不一致认识，所以需要采取谨慎的态度，更好地把好信息使用关。另外，信息的敏感程度不一样，分享的范围也不一样。更好的、精细的分享机制，应该是分层共享，即向征信系统报送什么数据，方能分享什么信息。

7) 征信服务宜由独立于信用交易当事人的第三方提供

征信业因其独特的行业特点，要求其保持很高的公信力。因此，虽然个别大的商业银行也有能力开展此项工作，但为了避免利益冲突，只能由独立于信贷业务之外的专业征信机构来做这项工作。征信机构本身是不能直接从事授信业务的。

征信机构的独立性，还体现在对原始数据的独立性上，即征信系统存储的关于各个信

息主体的原始数据，都是数据报送机构报送到系统的，征信机构无权修改，即便信用报告被确认有错误，也只能按流程由报送机构自己或由其授权，才能更正原始数据的错误。这种独立性是保证征信机构公信力的必要机制。

8) 征信的功能是在一定程度上揭示信息主体的信用状况

其采集的信息作为参考信息，可协助授信人或投资人更好地做出授信或投资决策。信用报告中的原始信用交易信息及征信增值产品如评分，都是用来帮助授信机构预测授信人未来的违约率，以帮助授信机构改善信用风险管理。

这里所说的“一定程度上”意思是指不同深度的征信活动，如基础征信或信用报告、信用调查、信用评级等业务活动，揭示信用状况的程度是不同的。

需要特别注意的是，信用报告和信用综合评价产品(如评级、评分等)并不是对评价对象诚信道德的评判，尽管诚信道德自身会对履约、信用状况产生一定影响。如果把征信活动与对信息主体的诚信道德评价等同起来，则有把征信业引向歧途的危险。实际上，世界上还没有哪一个国家有自动化的信息系统，试图对信息主体的道德进行评价。对一个人的道德评价是非常综合和复杂的。

9) 发展征信业应更多引入市场机制

征信活动主要是市场经济的产物，其运作也宜更多引入市场机制。

一是征信作为一种微观服务，是为了满足商业活动的需要。尽管征信机构也为政府部门等公共部门提供服务，但我们不能因此而把征信服务归为公共产品，征信主要是为市场化的授信机构服务的。已经实现市场化运作的授信机构，对征信服务的需求随市场变化而变化，这就要求征信业以十分灵活的机制来满足。二是征信专业服务的特征决定了由政府主导的公共机构不能充分发挥其功能。尽管世界上仍然存在一些由公共机构直接运作的征信系统，但其趋势是日渐式微的。

目前，国际上处于征信业垄断地位的个人和企业征信局均是市场化的征信机构，都是采取市场化的运作方式。这是保持较高的服务效率的需要。

10) 征信行业是依赖于法律法规，并受社会文化环境影响较大的新兴行业

征信数据的采集需要法律的支持，才能更好地进行。征信产品可以在什么范围使用，同样需要法律明确界定，才会避免争论。如何在促进征信体系发展与保护信息主体权益之间取得适当平衡，是征信立法需要解决的主要问题。

目前，我国征信业务还处于发展的初级阶段，有关法规还较少，只有中国人民银行2005 年发布的一个“3 号令”，即《个人信用信息基础数据库管理暂行办法》。因文化环境、道德理念等的不同，征信活动的法规在不同类型国家呈现出很大的差异性。如美国，有隐私法和公平信用报告法；英国，只有数据保护法；澳大利亚有数据保护法和征信机构行为守则；印度和俄罗斯等新兴市场经济国家，也有专门的规范征信机构的法规。

5. 征信的作用

征信活动服务的范围很广，如金融业、电信业、公共事业、政府部门等，从这些服务对象的不同角度出发，可以总结出征信具有以下 6 个作用。

1) 防范信用风险，促进信贷市场发展

随机波动理论认为，股价波动遵循随机波动，呈现典型的马尔可夫性质，股价过去的历史和从过去到现在的演变方式与股价的未来变动不相关。但是，对于单一个体而言，人类行为在很大程度上则具有路径依赖的特点，预测一个人未来行为的最好方法是看其过去的表现，这一点成为社会信用体系建设的理论基础。

银行如果不了解企业和个人的信用状况，为了防范风险，就会采取相对紧缩的信贷政策。通过征信活动，查阅被征信人以前的历史记录，商业银行能够比较方便地了解企业和个人的信用状况，采取相对灵活的信贷政策，扩大信贷范围，特别是对缺少抵押品的中小企业、中低收入者等边缘借款人。

2) 服务其他授信市场，提高履约水平

现代经济的核心是信用经济。授信市场包含的范围非常广泛，除银行信贷外，还包括大量的授信活动，如企业和企业(多以应收账款形式存在)、企业和个人(各种购物卡、消费卡等)、个人与个人(借款)之间的授信活动，一些从事授信中介活动的机构，如担保公司、租赁公司、保险公司、电信公司等，在开展业务时均需要了解受信方的信用状况。

征信活动通过信息共享、各种风险评估等手段将受信方的信息全面、准确、及时地传递给授信方，有效揭示受信方的信用状况，采用的手段有信用报告、信用评分、资信评级等。

3) 加强金融监管和宏观调控，维护金融稳定

通过征信机构强大的征信数据库，收录工商登记、信贷记录、纳税记录、合同履约、民事司法判决、产品质量、身份证明等多方面的信息，以综合反映企业或个人的信用状况。当从更为宏观的角度进行数据分析时，则可以整合出一个企业集团、一个行业和国家整体的信用风险状况。因此，可以按照不同的监管和调控需要，对信贷市场、宏观经济的运行状况进行全面、深入的统计和分析，统计出不同地区、不同金融机构、不同行业和各类机构、人群的负债、坏账水平等，从而为加强金融监管和宏观调控创造条件。

征信对监管者的帮助主要有两个：①监控总体信贷质量、测试银行是否满足监管要求(尤其是满足新巴塞尔资本协议要求)。例如，意大利的监管机构就利用征信数据库来测算商业银行的资本金要求、总体风险构成等，作为对商业银行进行监管依据的外部补充。②征信对宏观调控者的帮助主要体现在通过整体违约率的测算来判断经济目前所处的周期。

4) 服务其他政府部门，提升执法效率

征信机构在信息采集中除了采集银行信贷信息外，还依据各国政府的政府信息公开的法规采集了大量的非银行信息，用于帮助授信机构的风险防范。在这种情况下，当政府部门由于执法需要征信机构提供帮助时，可以依法查询征信机构的数据库，或要求征信机构提供相应的数据。

通过征信活动，使政府在依法行政过程中存在的信息不对称问题得到有效解决，为政府部门决策提供了重要的依据，这些依据主要是通过第三方反映出来的，信息的准确性比较强，有效地提高了执法效率。

5) 有效揭示风险，为市场参与各方提供决策依据

征信机构不仅通过信用报告实现信息共享，而且，会在这些客观数据的基础上通过加

工而推出对企业和个人的综合评价，如信用评分等。通过这些评价，可以有效反映企业和个人的实际风险水平，有效降低授信市场参与各方的信息不对称，从而得到市场的广泛认可，并帮助各方做出更好的决策。

这些综合评价主要有两个作用：第一是信号传递作用。通过这些综合评价，将新信息或现有的信息加以综合，提供给市场，市场根据这些综合评价所处的信用区间，对受信方的信用状况做出一个整体的评价。第二是证明作用。满足一定门槛的信用评分，往往成为监管者规定取得授信的条件之一。

6) 提高社会信用意识，维护社会稳定

在现代市场经济中，培养企业和个人具有良好的社会信用意识，有利于提升宏观经济运行效率。但是，良好的社会信用意识并不是仅仅依靠教育和道德的约束就能够建立的，必须在制度建设上有完备的约束机制。以美国为例，美国国民的社会信用意识和遵纪守法意识比较强，主要是靠完善的制度约束达致的，当制度约束缺失时，国民的社会信用意识和遵纪守法意识也会面临严峻的挑战。

征信在维护社会稳定方面也发挥着重要的作用。实践经验表明，不少企业和个人具有过度负债的冲动，如果不加约束，可能会造成企业和个人债务负担过重，影响企业和个人的正常经营和活动，甚至引发社会问题。有的国家就曾发生过信用卡过度发展，几乎酿成全民债务危机。一些西方国家建立公共征信机构的目的之一就是防止企业、个人过度负债，维护社会稳定。在我国，征信活动有助于金融机构全面了解企业和个人的整体负债状况，从制度上防止企业和个人过度负债，有助于政府部门及时了解社会的信用状况变动，防范突发事件对国计民生造成重大影响，维护社会稳定。

综上所述，正是因为征信能够帮助实现信息共享，提高对交易对手风险的识别，所以，征信在经济和金融活动中具有重要的地位，构成了现代金融体系运行的基石，是金融稳定的基础，对于建设良好的社会信用环境具有非常深远的意义。

7.1.2 征信的基本流程

征信活动可以分为两类：一类是征信机构主动去调查被征信人的信用状况；另一类是依靠授信机构或其他机构批量报送被征信人的信用状况。两者最大的区别在于前者往往是一种个体活动，通过接受客户的委托，亲自到一线去收集调查客户的信用状况，后者往往是商业银行等授信机构组织起来，将信息定期报给征信机构，从而建立信息共享机制。两者还有一个区别是前者评价的范围更广，把被征信人的资质情况、诚信度考察、资产状况等都包括在内，而后者由于是批量采集信息，因此灵活性和主观性上不如前者，但规律性和客观性则强于前者。但两类方式在征信的基本流程上是相同的，例如，前一类流程要制订计划，决定采集哪些信息，而后一类流程也同样如此，由征信机构事先确定好需要采集的信息后，与信息拥有方协商，达成协议或其他形式的约定，定期向征信机构批量报送数据，因此，在讨论流程时，可以将两者合并在一起，如图 7.2 所示。

图 7.2　征信数据库形成流程

1. 制订数据采集计划

能够反映被征信人信用状况的信息范围广泛，为提高效率、节省成本，征信机构应事先制订数据采集计划，做到有的放矢。这是征信基本流程中一个重要的环节，一份好的计划能够有效减轻后面环节的工作负担。一般来说，数据采集计划包括以下内容。

1)　采集数据项

客户使用征信产品的目的都不尽相同，有的希望了解被征信人短期的信用状况，有的则是作为中长期商业决策的参考。客户的不同需求决定了数据采集重点的迥异。征信机构要本着重点突出、不重不漏的原则，从客户的实际需求出发，进而确定所需采集数据的种类。例如，A 银行决定是否对 B 企业发放一笔短期贷款时，应重点关注该企业的历史信贷记录、资金周转情况，需采集的数据项为企业基本概况、历史信贷记录、财务状况等。

2)　采集方式

确定科学合理的采集方式是采集计划的另一主要内容。不论主动调查，还是授信机构或其他机构批量报送数据，征信机构都应制定最经济便捷的采集方式，做好时间、空间各项准备工作。对于批量报送数据的方式，由于所提供的数据项种类多、信息量大，征信机构应事先制订一个规范的数据报送格式，让授信机构或其他机构按照格式报送数据。

3)　其他事项

在实际征信过程中，如果存在各种特殊情况或发生突发状况，征信机构应在数据采集计划中加以说明，以便顺利开展下面的工作。

2. 采集数据

数据采集计划完成后，征信机构应依照计划开展采集数据工作。数据一般来源于已公开信息、征信机构内部存档资料、授信机构等专业机构提供的信息、被征信人主动提供的信息、征信机构正面或侧面了解到的信息。出于采集数据真实性和全面性的考虑，征信机构可通过多种途径采集信息。但要注意，这并不意味着数据是越多越好，要兼顾数据的可用性和规模，在适度的范围内采集合适的数据。

3. 数据分析

征信机构收集到的原始数据，只有经过一系列科学分析之后，才能成为具有参考价值的征信数据。

1) 数据查证

数据查证是保证征信产品真实性的关键步骤。一查数据的真实性。对于存疑的数据，征信机构可以通过比较不同采集渠道的数据，来确认正确的数据。当数据来源唯一时，可通过二次调查或实地调查，进一步确定数据的真实性。二查数据来源的可信度。某些被征信人为达到不正当目的，可能向征信机构提供虚假的信息。如果发现这种情况，征信机构除及时修改数据外，还应记录该被征信人的“不诚信行为”，作为以后业务的参考依据。三查缺失的数据。如果发现采集信息不完整，征信机构可以依据其他信息进行合理推断，从而将缺失部分补充完整。比如，利用某企业连续几年的财务报表推算出某几个数据缺失项。最后是被征信人自查，即异议处理程序。当被征信人发现自己的信用信息有误时，可向征信机构提出申请，修正错误的信息或添加异议声明。特别是批量报送数据时，征信机构无法对数据进行一一查证，一般常用异议处理方式。

2) 信用评分

信用评分是个人征信活动中最核心的数据分析手段，它运用先进的数据挖掘技术和统计分析方法，通过对个人的基本概况、信用历史记录、行为记录、交易记录等大量数据进行系统的分析，挖掘数据中蕴含的行为模式和信用特征，捕捉历史信息和未来信息表现之间的关系，以信用评分的形式对个人未来的某种信用表现做出综合评估。信用评分模型有各种类型，能够预测未来不同的信用表现。常见的有信用局风险评分、信用局破产评分、征信局收益评分、申请风险评分、交易欺诈评分、申请欺诈评分等。

3) 其他数据分析方法

在对征信数据进行分析时，还有许多其他的方法，主要是借助统计分析方法对征信数据进行全方位分析，并将分析获得的综合信息用于不同的目的，如市场营销、决策支持、宏观分析、行业分析等领域。使用的统计方法主要有关联分析、分类分析、预测分析、时间序列分析、神经网络分析等。

4. 形成信用报告

征信机构完成数据采集后，根据收集到的数据和分析结果，加以综合整理，最终形成信用报告。信用报告是征信机构前期工作的智慧结晶，体现了征信机构的业务水平，同时也是客户了解被征信人信用状况、制定商业决策的重要参考。因此，征信机构在生成信用

报告时，务必要贯彻客观性、全面性、隐私和商业秘密保护的科学原则。所谓客观性，指的是信用报告的内容完全是真实客观的，没有掺杂征信机构的任何主观判断。基于全面性原则，征信报告应充分披露任何能够体现被征信人信用状况的信息。但这并不等于长篇大论，一份高质量的信用报告言简意赅、重点突出，使客户能够一目了然。征信机构在撰写信用报告过程中，一定要严格遵守隐私和商业秘密保护原则，避免泄露相关信息，致使客户和被征信人权益受到损害。信用报告是征信机构最基本的终端产品。随着征信技术的不断发展，征信机构在信用报告的基础上衍生出越来越多的征信增值产品，如信用评分等。不论形式如何变化，这些基本原则是始终不变的。

7.1.3 征信行业产业链

征信行业产业链包括上游的数据生产者、中游的征信机构及下游的征信信息使用者，其中中游的征信机构运行模式主要有采集数据、加工数据及销售产品。征信行业产业链如图 7.3 所示。

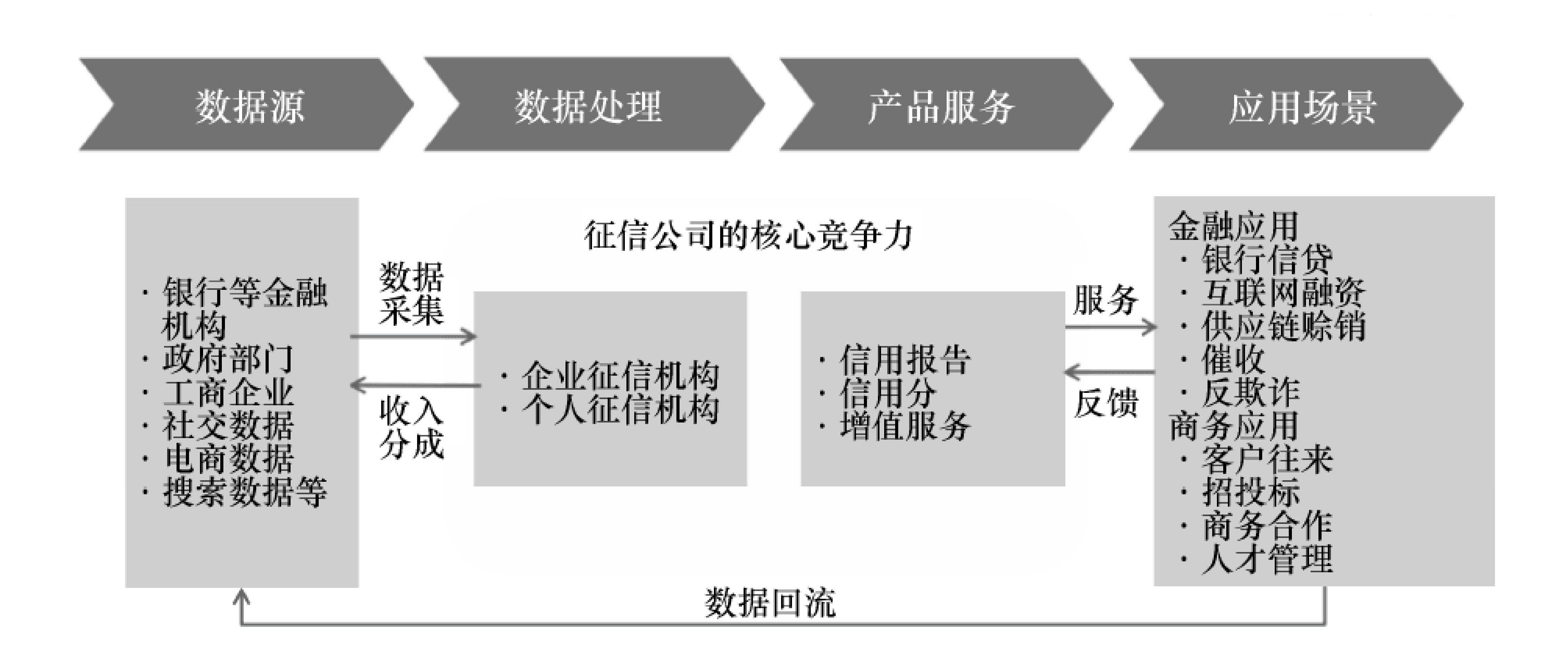


图 7.3 征信行业产业链划分

按照数据生产者划分，可以分为个人征信和企业征信。个人征信的数据生产者是个人，征信机构采集个人产生的数据，加工并销售信用产品；而企业征信的数据生产者是企业(工商企业、政府、金融机构或是小微企业等)，征信机构采集企业生产的数据，加工及销售信用产品。

7.1.4 征信产品

1. 企业征信产品

经过多年的探索和发展，企业征信系统的产品和服务体系日益完备，以各种版本信用报告为核心的基础产品体系已经相对成熟，以关联查询服务、企业征信汇总数据为代表的增值服务体系初步形成。

1) 基础产品

企业信用报告是企业征信系统提供的基础产品。随着征信系统应用的推广与深入，信

用报告已成为商业银行信用风险管理的重要工具，服务于银行信贷流程中的贷前审查、贷后管理、资产保全等各个环节。

改进企业信用报告版本。2005 年，征信中心首次推出企业信用报告时，仅有一个版本。为更好地服务不同类别的用户，征信中心不断优化信用报告内容、丰富信用报告版本。新版企业信用报告于 2013 年正式推出。

新版企业信用报告针对不同的需求主体分为 4 个版本：一是为以银行为代表的授信机构服务的银行版；二是为政府部门履职使用的政府版；三是为其他机构服务的社会版；四是为满足信息主体查询需求的自主查询版。新版信用报告内容更加丰富、完整；结构层次更分明，信息展示顺序更加符合阅读习惯，展示方式更加灵活，可读性更强。

企业信用报告的主要内容包括报告头、基本信息、有直接关联关系的其他企业、财务报表、信息概要、信贷记录明细、公共记录明细、声明信息明细等。不同版本的企业信用报告，内容各有侧重。新版企业信用报告的基本内容如图 7.4 所示。

	银行版	政府版	社会版	自主查询版
报告头	√	√	√	√
报告说明	√	√	√	√
基本信息	√	√	√	√
有直接关联关系的其他企业	√	无	无	√
财务报表	√	无	无	无
信息概要	√	无	√	√
信贷记录明细	√	无	无	√
公共记录明细	√	√	√	√
声明信息明细	√	√	√	√

图 7.4 新版企业信用报告的基本内容

目前，根据服务对象和使用目的的不同，各类用户可以通过页面方式和接口方式查询企业信用报告。

2) 增值产品

征信中心对所采集的各类企业信息进行深加工，针对用户的个性化需求，先后推出了关联企业查询、企业征信汇总数据、对公业务重要信息提示、征信系统信贷资产结构分析、历史违约率等增值产品。

(1) 关联企业查询。

关联企业查询产品是基于企业征信系统借款人基本信息和信贷信息，通过数据挖掘找出借款人与企业、借款人与个人存在的直接或间接或共同控制的经济关系，包括以资本为纽带和以经济利益为纽带的 33 种关系。

目前，征信中心主要提供 3 类关联企业查询产品：一是关联企业名单及关系表；二是关联企业群信贷业务及被起诉信息汇总表；三是关联企业群的贷款业务集中还款时间统计表。

早在 2002 年，银行信贷登记咨询系统就开始提供这项服务。企业征信系统上线以

后，经过多轮改造，关联关系达到 9 大类 33 种；提供方式由标准化转为个性化；查询方式在单个查询基础上增加了批量查询；服务模式由来函申请查询转为在线查询；产品由标准化转为自定义；服务对象由单一自身使用扩大到 10 个政府部门、各省人民银行分支行和银监局、21 家全国性商业银行和 580 多家地方性金融机构；应用面由单一集团客户信贷管理，扩大到小微企业信贷管理。

(2) 企业征信汇总数据。

该产品是利用企业征信系统的数据，以金融统计核算原则为基础，通过对数据的加工、整理，建立银行业信贷业务报表体系和指标体系，综合反映银行业信贷业务的运行特征和状况，从而为货币政策制定、金融监管和商业银行经营管理提供全面、及时、准确的信息。其服务对象主要是人民银行各级分支机构。

2007—2011 年，主要以来函申请的方式提供查询。2011 年，征信中心建成征信数据应用分析系统，实现在线查询，服务效率提高。此后，随着该系统业务处理流程的优化，服务时效性进一步增强。

企业征信汇总数据主要包括信贷结构类汇总数据和信贷特征类汇总数据两类。前者于 2007 年 10 月正式投产使用，目前按月向各征信分中心提供辖内信贷汇总数据，也为人民银行及其分支机构的个性化需求提供服务。后者于 2011 年 12 月正式上线，主要服务于各人民银行分支机构，用于为本辖区的货币政策执行和金融风险监控提供信息参考。

(3) 对公业务重要信息提示。

该产品是利用企业征信系统即时更新的数据，每工作日将各机构用户的本机构“好客户”在其他机构发生“新增逾期 90 天/60 天”、五级分类“新增不良”“新增失信被执行人”等提示信息主动推送给相关机构用户总部。

(4) 征信系统信贷资产结构分析。

该产品是运用征信系统的数据，以图形的形式反映单家机构在信贷市场中的相对位置以及市场份额，为商业银行信贷决策提供信息支持。该产品指标设计以行业、地区为主线，以贷款、贸易融资、票据贴现、保理、信用证、银行承兑汇票、保函等 7 项业务为辅线，提供分地区、分行业、分信贷品种的信贷市场运行分析、信贷市场结构分析、信贷资产质量分析。每类指标既提供时点(或时段)值，又提供时间序列值，均以图形的形式展示。

(5) 历史违约率。

该产品利用征信系统覆盖全市场的数据计算出某一时点上的正常客户，之后 1 年在全市场上发生违约的比率。该产品包括客户在本银行和他银行的违约，反映银行业对公业务中借款人平均违约水平，可作为行业中衡量这一群体实际违约水平的标准，直接用于校准商业银行使用本银行数据计算的历史违约比率，提高测算违约概率的精准度，为商业银行配置信贷资产组合和定价、制订信贷方案提供数据支持。

历史违约率产品分两大类：一是银行业所有客户的违约率；二是本机构客户在银行业发生信贷业务的违约率。该产品按月加工，向用户提供分行业、地区(借款人注册地和金融机构所在地)、借款人规模、金融机构(全金融机构和本机构)、信贷业务种类、违约标准 6 个查询条件。查询结果包括期初正常客户数、观察期违约客户数、违约率值。

2. 个人征信产品

经过 10 年的积极探索和经验积累，个人征信系统已形成以个人信用报告、个人信用信息提示和个人信用信息概要为核心的基础产品体系；以个人业务重要信息提示和个人信用报告数字解读为代表的增值产品体系。

1) 基础产品

个人征信系统提供的基础产品主要有个人信用报告、个人信用信息提示和个人信用信息概要 3 种。

(1) 个人信用报告。

个人信用报告是个人征信系统提供的核心基础产品。多年来，征信中心通过不断优化个人信用报告内容、丰富信用报告版本、完善信用报告版式设计等方式，促进个人信用报告更好的应用。

目前，个人信用报告根据服务对象及使用目的不同，分为 4 个版本：为以银行为代表的授信机构服务的银行版，含配套的仅包含本行报送信息的银行异议版；满足消费者本人查询需求的个人版(含彩色样式)以及个人明细版(彩色样式)；为其他社会主体服务的社会版；供征信系统管理使用的征信中心版。个人信用报告的基本内容包括：报告头、个人基本信息、信贷交易信息、公共信息、声明信息、查询记录和报告说明。不同版本的信用报告对上述内容各有侧重。新版个人信用报告的主要内容如图 7.5 所示。

报告内容	银行版	银行异议处理版	个人版（含彩色样式）	个人版（明细）	征信中心版	社会版
报告头	√	√	√	√	√	√
基本信息	√	√	√	√	√	无
信息概要	√	无	√	√	√	√
信贷交易信息明细	√	√	√	√	√	无
公共信息	√	无	√	√	√	√
声明信息	√	√	√	√	√	√
查询记录	√	√	√	√	√	√
报告说明	√	√	√	√	√	√
备注	屏蔽他行的机构名称和业务号	仅包含本机构报送的信贷信息	基本信息仅包含婚姻状况			

图 7.5 新版个人信用报告的主要内容

(2) 个人信用信息提示。

信用信息提示是用来提示个人信息主体在个人征信系统中是否存在最近 5 年的逾期记录，通过互联网个人信用信息服务平台和短信方式向个人信息主体提供查询服务。

(3) 个人信用信息概要。

个人信用信息概要主要包括信贷记录、公共记录和最近 2 年内查询记录的汇总统计信息，便于消费者快速了解自身的信用概况，通过互联网个人信用信息服务平台向信息主体提供查询服务。

2) 增值产品

(1) 个人业务重要信息提示。

个人业务重要信息提示是利用个人征信系统即时更新的数据，按周将各机构用户的本机构“好客户”在其他机构发生“新增逾期 61～90 天/90 天以上”、贷款五级分类“新增

不良”、信用卡账户状态“新增呆账”、贷款或信用卡“新增账户”。

“新增失信被执行人”等提示信息主动推送给相关机构用户总部。信息提示方式包括页面展示和下载、接口主动推送、邮件主动推送 3 种，用户可自行选择使用。

需要引起注意的是，个人业务重要信息提示不同于个人信用信息提示。两者的主要区别是：个人业务重要信息提示是面向授信机构用户提供的服务；而个人信用信息提示是面向个人信息主体提供的服务。

(2) 个人信用报告数字解读。

个人信用报告数字解读(以下简称“数字解读”)是在征信中心与美国费埃哲公司(Fair Isaac Corporation)合作进行个人征信评分研究项目的基础上，利用个人征信系统的信贷数据，使用统计建模技术开发出来的个人信用风险量化服务工具，用于预测放贷机构个人客户在未来一段时间内发生信贷违约的可能性，并以“数字解读”值的形式展示。

“数字解读”的分数范围为 0～1000 分，每个分数对应一定的违约率。分值越高，表示未来发生信贷违约的可能性越低，其信用风险越小；分值越低，表示未来发生信贷违约的可能性越高，其信用风险越大。一般情况下，高分人群整体的信用状况优于低分人群，即未来发生信贷违约的可能性较低。“数字解读”旨在帮助放贷机构更加便捷地使用信用报告信息，了解客户的信贷风险状况及未来发生信贷违约的可能性。

7.1.5 征信机构

征信机构是负责管理信用信息共享的机构，从事个人和(或)企业信用信息的采集、加工处理，并为用户提供信用报告和其他基于征信系统数据的增值产品。

从全球实践来看，征信机构一般分为 3 类：个人征信机构(credit bureau)、信贷登记系统(credit registry)和企业征信机构(commercial credit reporting company)，3 类机构的经营模式和目标服务市场各有差异。

1. 个人征信机构

个人征信机构(credit bureau)通常是私营的，是按照现代企业制度建立、完全市场化运作的征信机构，主要为商业银行、保险公司、贸易、邮购公司等信息使用者提供服务。美国是典型的私营征信机构模式，商业化征信机构拥有全面的信用信息系统。

个人征信机构主要为信贷机构提供个人借款人以及微型、中小型企业的信用信息。它们从银行、信用卡发行机构和其他非银行金融机构等各类信贷机构采集标准化的信息，同时还采集各类公共信息，如法院判决、破产信息、电话簿信息，或担保物权登记系统等第三方数据库的信息。此外，它们也会采集一些非传统信用数据，如零售商对消费者的赊销信息，以及煤气、水、电等公共事业缴费信息，有线电视、电话、网络等其他先使用服务后付费服务的缴费数据，以便提供更好、更完善的信用报告。对从未与银行发生过信贷关系的个人以及微型、中小型企业而言，不断拓宽信息来源非常有益，可以帮助它们在没有银行信贷记录的情况下建立起信用档案，从而有效解决因为没有信用档案而无法获得银行贷款的难题。

一直以来，个人征信机构主要采集个人信息。近年来，随着微型以及中小企业信贷业

务的发展、信息技术的进步，越来越多的个人征信机构开始采集微型以及中小企业的信用信息，并提供其信用报告。根据世界银行《2012 全球营商环境报告》对全球 100 家个人征信机构的调查，超过 80%或多或少都采集企业信息。这样做的好处是可以把对企业与业主的信用评估结合起来，因为微型和中小企业的业主经常把个人财务和企业财务混在一起，所以企业业主的信用记录是评估小企业信用风险的重要参考因素。

个人征信机构通常采取数据提供者自愿报数(通过签署数据共享互惠协议)的模式，广泛采集各类信用数据，并提供多样化的征信产品和服务，帮助信贷机构做出信贷决策。在一些国家和地区，通常是在征信业的发展初期，法律会强制要求有关各方进行数据共享，并使用征信机构的服务。此外，还会赋予监管机构相应的权利，以督促信贷机构加入征信系统并监控其加入情况。

2. 信贷登记系统

信贷登记系统起源于欧洲。从历史上看，信贷登记系统(credit registry)的建立目的与个人征信机构不同。大多数信贷登记系统最初是作为中央银行的内部数据库而设立的，而且目前仍然有很多信贷登记系统用于中央银行的宏观金融监管。根据世界银行的调查，越来越多的国家政府鼓励成立信贷登记系统来监督商业银行的信贷活动。因此，这些数据库通常采集贷款额度在一定金额以上的大额信贷业务数据。最初，信贷登记系统的信息仅限于央行内部使用。但随着时间的推移，信贷登记系统也开始向受监管的信贷机构提供信用报告。而且，随着消费信贷的发展，信贷登记系统普遍降低或取消了数据采集门槛。在许多国家，如法国、阿根廷、西班牙、秘鲁、意大利、比利时等，信贷登记系统已经开始提供与个人征信机构类似的产品和服务。通常，法律要求所有受监管的金融机构都要向信贷登记系统报送数据。

信贷登记系统既采集个人信息，也采集企业信息。个人信息通常包括个人的身份验证信息、贷款类型和贷款特征信息、负面信息、担保和保证类信息以及还款记录信息。企业信息通常包括企业的身份标识信息、企业主的信息、贷款类型和贷款特征信息、负面信息和还款记录。

3. 企业征信机构

企业征信机构(commercial credit reporting company)提供关于企业的信息，这些企业包含个人独资企业、合伙企业和公司制企业，并通过公共渠道、直接调查、供货商和贸易债权人提供的付款历史来获取信息。企业征信机构所覆盖的企业在规模和经营收入上都小于信用评级机构所覆盖的企业，其采集的信息一般用于信用风险评估或信用评分，或是用于贸易信用展期等其他用途。

企业征信机构与个人征信机构的差异体现在以下几个方面：企业征信机构采集的信息不包括个人敏感信息，所覆盖的交易的规模也大得多。与个人征信相比，企业征信往往需要采集更多的有关企业借款人的支付信息和财务信息。为了保护个人数据主体的权利，个人征信机构会披露数据提供者的身份，但企业征信机构却不会让企业数据主体知道其数据来源或用户的身份。

企业征信机构也可能会采集小企业的信息，但由于其报告的数据项并不适合小企业，

所以采集的信息往往比较有限。正如前面提到的，由于小企业往往不会公开自身的财务信息，所以其企业主的信用记录对评估小企业的信用情况非常有用。但企业征信机构并不采集个人数据。此外，由于微型或小型企业的信用信息采集成本往往较高。因此，与企业征信机构相比，个人征信机构往往能更好地满足对微型和中小型企业的征信需求。

除以上征信机构外，现实中，很多发展中国家的征信机构以欧洲模式为基础，在发展过程中又向美国模式倾斜，呈现出混合模式的特点。

国际上征信系统的建设和运行越来越呈现出多元化特别是运行市场化的特点，很难说某个征信机构是公共的还是纯商业性的。机构的性质不只是两个极端，即便政府出资举办的事业或投资的机构，很多也是在不同程度上进行市场化运作的。目前新兴市场经济国家尤其是亚洲地区的征信机构都是运用市场的力量来建设和运行公共征信系统，针对为金融机构和社会提供的服务收费，实现征信机构的可持续发展。

7.1.6 征信体系

征信体系是指与征信活动有关的法律规章、组织机构、市场管理、文化建设、宣传教育等共同构成的一个体系。

征信体系的主要功能是为信贷市场服务，但同时具有较强的外延性，还向商品交易市场和劳动力市场提供服务。在实践中，征信体系的主要参与者有征信机构、金融机构、企业、个人及政府。社会信用体系是市场经济发展的必然产物。在信用交易成为市场交易的主要方式、信用工具被大规模使用以及信用风险日益显著的背景下，社会信用体系成为影响一个国家经济发展的重要方面。经过上百年的市场经济发展，发达国家形成了相对比较完善的社会信用体系。但是，由于各国经济、文化、历史不同，不同国家形成了不同的社会信用体系模式。

1. 国外征信体系模式

如图 7.6 所示为世界征信体系模式的种类。

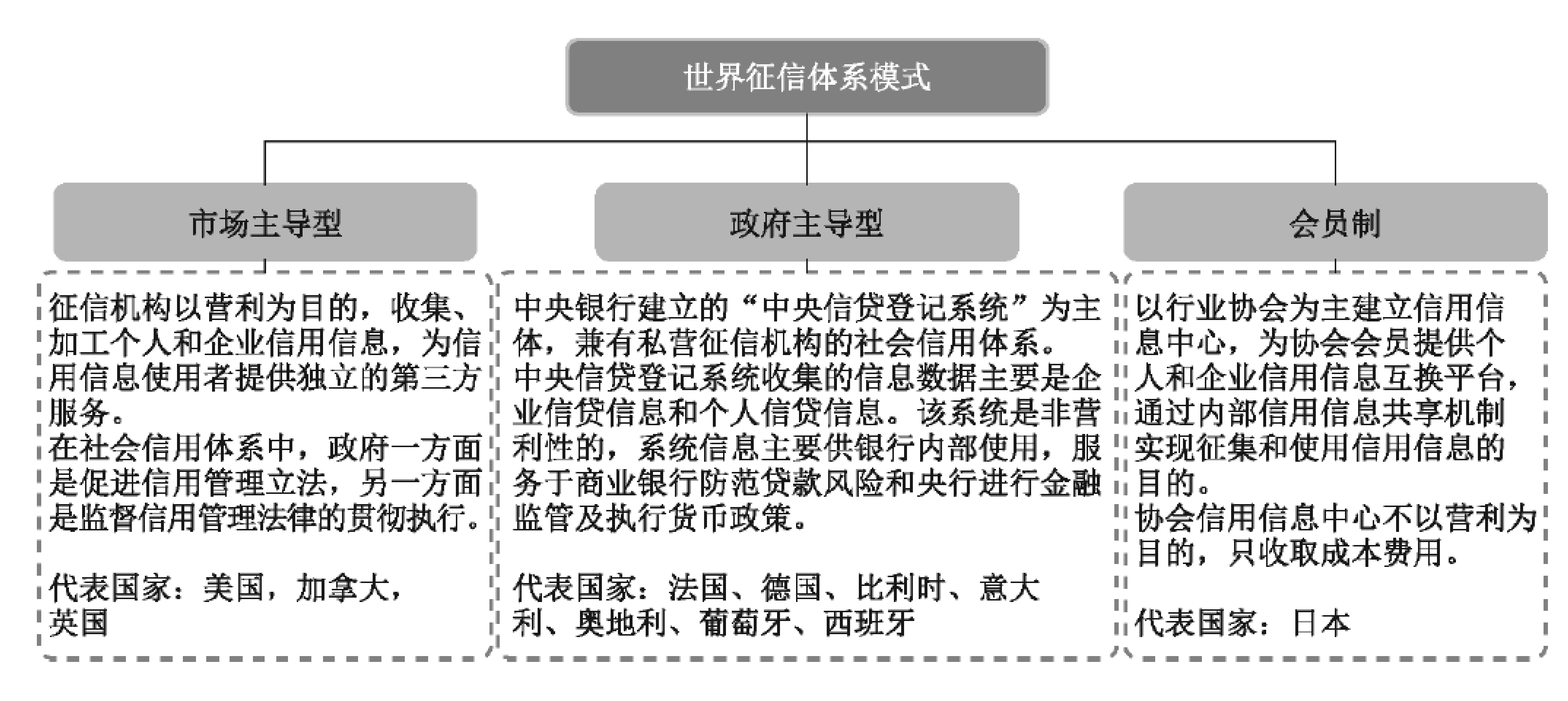


图 7.6 世界征信体系模式的种类

从国际发达国家的经验看，征信体系模式主要有 3 种。

第一种，市场主导型模式。又称民营模式。这种社会信用体系模式的特征是征信机构以营利为目的，收集、加工个人和企业的信用信息，为信用信息的使用者提供独立的第三方服务。在社会信用体系中，政府的作用一方面是促进信用管理立法，另一方面是监督信用管理法律的贯彻执行。美国、加拿大、英国和北欧国家采用这种社会信用体系模式。

第二种，政府主导型模式。又称公共模式或中央信贷登记模式。这种模式是以中央银行建立的“中央信贷登记系统”为主体，兼有私营征信机构的社会信用体系。中央信贷登记系统是由政府出资建立的全国数据库网络系统，直接隶属于中央银行。中央信贷登记系统收集的信息数据主要是企业信贷信息和个人信贷信息。该系统是非营利性的，系统信息主要供银行内部使用，服务于商业银行防范贷款风险和央行进行金融监管及执行货币政策。据世界银行统计，法国、德国、比利时、意大利、奥地利、葡萄牙和西班牙 7 个国家有公共信用登记机构，即中央信贷登记系统。其中，除法国外，其他 6 国都有市场化运营的私人征信机构。

第三种，会员制模式。它是指由行业协会为主建立信用信息中心，为协会会员提供个人和企业的信用信息互换平台，通过内部信用信息共享机制实现征集和使用信用信息的目的。在会员制模式下，会员向协会信息中心义务地提供由会员自身掌握的个人或者企业的信用信息，同时协会信用信息中心也仅限于向协会会员提供信用信息查询服务。这种协会信用信息中心不以营利为目的，只收取成本费用。日本采用这种社会信用体系模式。

1) 美国的市场主导型模式

美国的征信业始于 1841 年，第一家征信所是由纽约的一名纺织批发商刘易斯·塔潘所建立。1870 年，R. G. 邓恩接管了这家征信所，后来又与布雷兹特里特征信所合并，组成 Dun & Bradstreet。从简单征信服务到比较完善的现代信用体系的建立，美国的征信业差不多经历了 160 多年的时间。

“美国模式”是典型的市场主导型。美国的征信服务机构都是独立于政府之外的民营征信机构(或称为私人信用调查机构)，是按照现代企业制度方式建立，并依据市场化原则运作的征信服务主体。

美国的征信服务机构具有明显特征：①在机构组成方面，征信机构主要由私人和法人投资组成。②在信息来源方面，民营征信机构的信息来源广泛。消费者信用调查机构的信用信息除了来自银行和相关的金融机构外，还来自信贷协会和其他各类协会、财务公司或租赁公司、信用卡发行公司和商业零售机构等。③在信用信息内容方面，民营征信机构的信息较为全面，不仅征集负面信用信息，也征集正面信息。④在服务范围方面，美国民营信用调查机构是面向全社会提供信用信息服务。服务的对象主要包括私人银行、私人信用机构、其他企业、个人、税收征管机构、法律实施机构和其他联邦机构，以及本地政府机构等，这些机构都是征信报告的需求方。

美国对征信的立法是由于 20 世纪 70 年代征信业快速发展导致了一系列问题而开始，走的是一条在发展中规范的立法过程。到现在美国不仅具备了较为完善的信用法律体系和政府监管体系，而且与市场经济的发展相伴随，形成了独立、客观、公正的法律环境。政府基本上处于社会信用体系之外，主要负责立法、司法和执法，建立起一种协调的市场环

境和市场秩序，同时其本身也成为商业性征信公司的评级对象，这样就保证了征信公司能确保其独立性、中立性和公正性。

美国的信用管理法律制度可以划分为 3 个层次。第一层次是直接的信用管理法律规定。第二层次是直接保护个人隐私的法律，这些法律都直接规定，在相应的特殊环境中不能公布或者限制公布个人或企业的相关信息。第三层次是指规范政府信息公开的法律，为征信机构收集政府公开信息提供法律依据。

2) 欧洲的政府主导型模式

欧洲征信业的发展主要采用的是政府主导型模式。欧洲对于征信的立法最初是源于对数据、个人隐私的保护，因此与美国相比，欧洲具有较严格的个人数据保护法律。

欧洲的政府主导型征信模式与美国的市场化模式的差别体现在 3 个方面：一是信用信息服务机构是被作为中央银行的一个部门建立，而不是由私人部门发起设立。二是银行需要依法向信用信息局提供相关信用信息。三是中央银行承担主要的监管职能。

3) 日本的会员制征信模式

日本的征信体系明显区别于美国和西欧国家，采用的是会员制征信模式，这主要是由于日本的行业协会在日本经济中具有较大的影响力。尤其对于个人征信而言，在日本没有商业化运作的个人征信企业。

目前，日本的信用信息机构大体上可划分为银行体系、消费信贷体系和销售信用体系 3 类。相应的行业协会分别是银行业协会、信贷业协会和信用产业协会。这些协会的会员包括银行、信用卡公司、保证公司、其他金融机构、商业公司、零售店等。三大行业协会的信用信息服务基本能够满足会员对个人信用信息征集考查的需求。例如，日本银行协会建立了非营利的银行会员制机构，即日本个人信用信息中心，地方性的银行作为会员参加“信息中心”。到 1988 年，全国银行协会把日本国内的信息中心统一起来，建立了全国银行个人信息中心。信息中心的信息来源于会员银行，会员银行在与个人签订消费贷款的合同时，均要求个人义务提供真实的个人信用信息。这些个人信息中心负责对消费者个人或企业进行征信。该中心在收集与提供信息服务时要收费，以维持中心的运行与发展，但不以营利为目的。不过，日本征信业同时也存在一些商业性的征信公司。

日本的消费者信用信息并不完全公开，只是在协会成员之间交换使用，以前并无明确的法律规定，但在银行授信前，会要求借款人签订关于允许将其个人信息披露给其他银行的合同。日本也注重完善了有关保护个人隐私的基本法律，重点确定个人金融信用信息、医疗信息、通信信息的开放程度。

2. 我国征信体系模式

1) 我国的政府主导型模式

当前，政府主导型的征信机构占据绝对优势。外商独资型公司的服务对象主要是外商且规模较小。而中外合资的征信机构发展势头较快，私营征信机构发展受到的限制最大。

与美、德、日相对比，我国公共征信机构占主导地位，私人征信机构数量和规模都很小，发展前景广阔。根据国际经验，一国个人征信机构体系应与本国征信业的发展特点相适应，相较于美国的完全市场化模式和日本的协会模式，我国与欧洲的政府主导模式可能会更为相近。

2) 征信体系的框架构成

征信体系是伴随着信用经济的发展，逐步形成的相互联系的整体结构。它是客观存在的系统性体系结构，包含许多信用经济乃至市场经济发展过程中必备的子领域，共同构成信用经济发展不可或缺的市场服务和监督系统，保障信用经济的健康稳定发展，维护正常的信用经济秩序和环境。社会征信体系如图 7.7 所示。

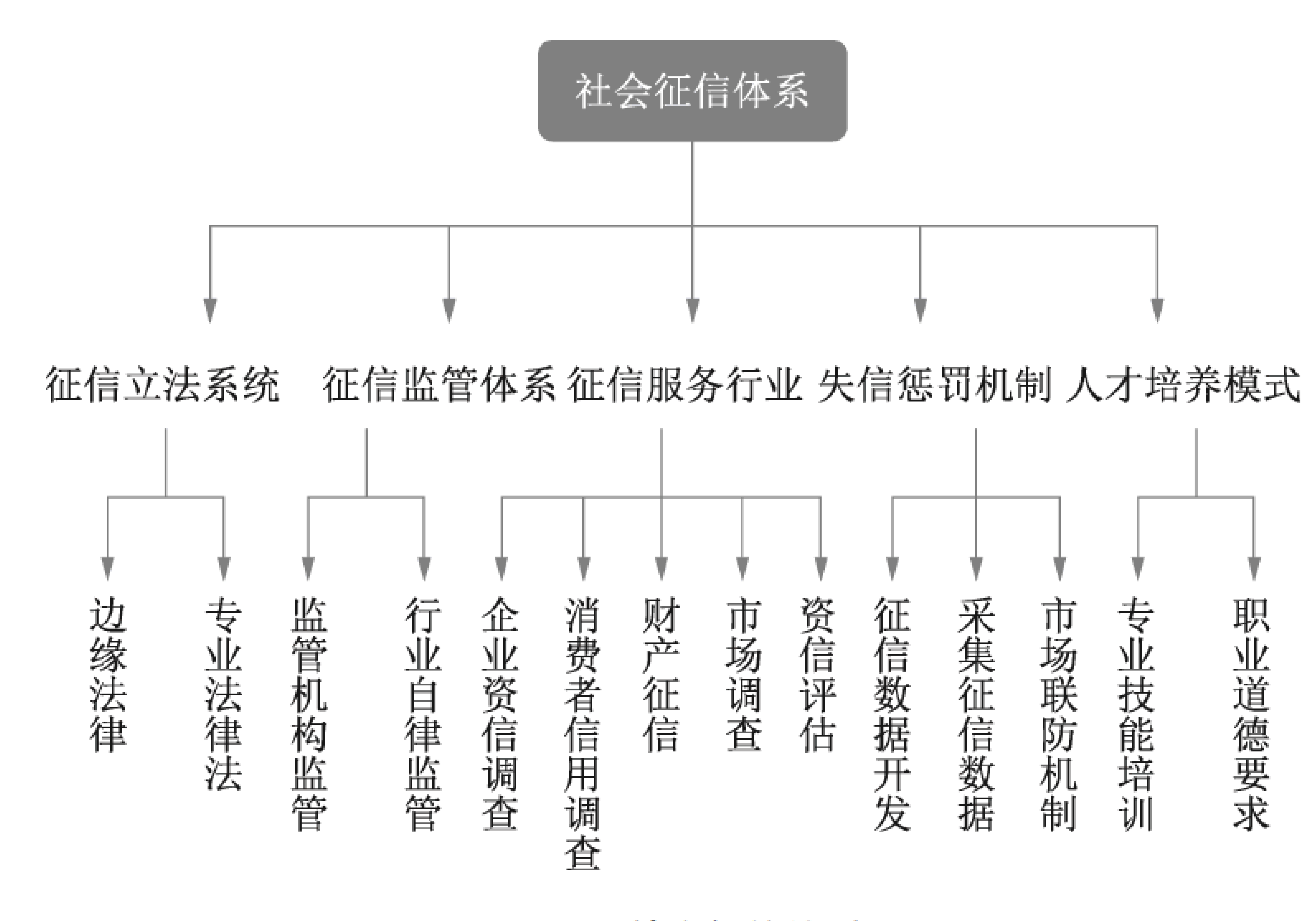


图 7.7 社会征信体系

一般来讲，征信体系包含一些相互联系的子体系，具体如下。

(1) 征信立法系统。征信立法系统主要负责征信行业法律法规的制定和执行，监管征信行业发展的规范性和维护征信市场的市场准则，同时督促征信行业内部进行行业自律，从外部和内部加强对征信行业发展的指导和监督。其中，不仅有与征信相关的基础法律体系(一般称之为“边缘法”)，而且也有征信专业法律系统。

(2) 征信监管体系。征信体系为社会提供征信服务，除了在法律允许范围内开展业务之外，还要有必要的监管体系来行使约束其行为的职能。征信监管体系主要由政府专设的监管机构和行业自律组织负责整个征信服务行业的管理和指导。

(3) 征信服务行业。在信用管理行业内部，征信服务是一种基础性服务。它受委托人的委托进行调查，以一种或若干种调查和分析报告类的征信产品作为回复，帮助委托人获取信用信息，以便做出合适的决策。

(4) 失信惩罚机制。所谓失信惩罚机制，它是社会征信体系中重要的“部件”之一，主要通过经济手段和道德谴责，惩罚市场经济活动中的失信者，将有严重经济失信行为的企业和个人从市场的主流中剔除出去。同时，失信惩罚机制可以使政策向诚实守信的企业和消费者倾斜，间接降低守信用企业获取资本和技术的门槛，消除障碍和壁垒。失信惩罚机制的最大特征就是对失信行为的出击是主动的，而不像征信服务那样是被动的，只有在接受委托人的具体委托后才提供征信服务。这样，失信惩罚机制和征信服务可以主动与被动相结合，共同维护征信体系的良好运行，保证健康有序的经济秩序。失信惩罚机制的工作原理如图 7.8 所示。

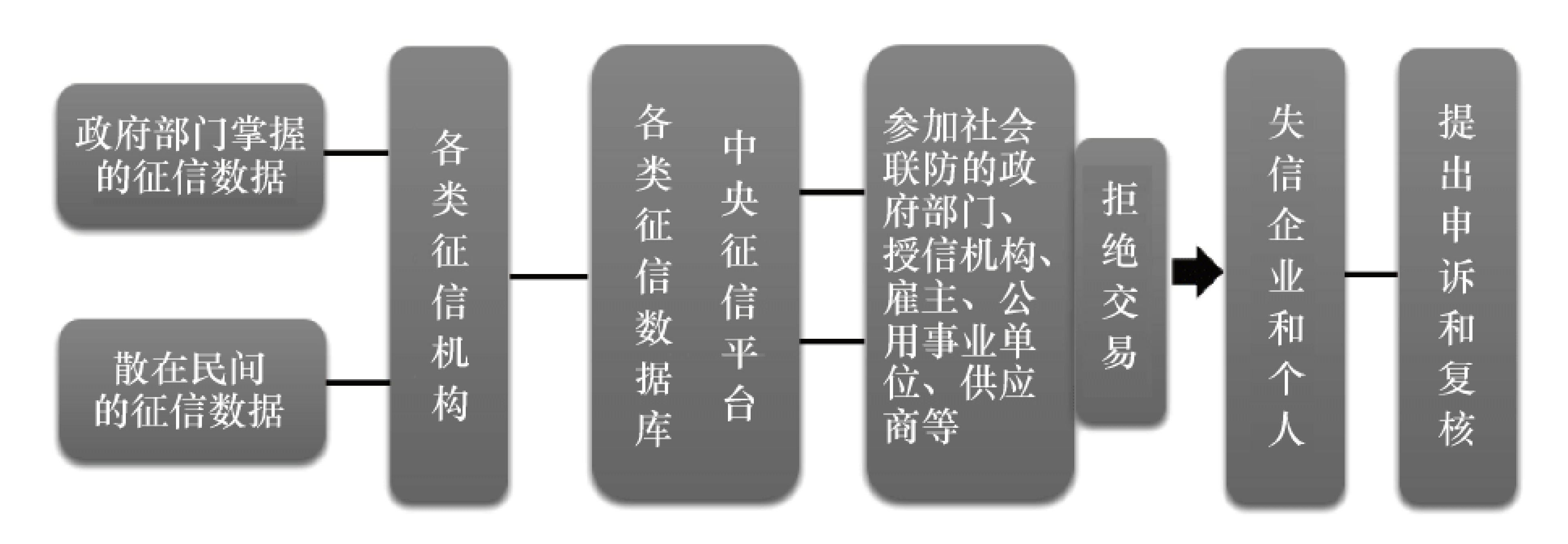


图 7.8　失信惩罚机制的工作原理

(5)　人才培养模式。征信行业是一个对于经验要求比较高、知识面要求比较广的行业。在这一领域的从业人员，除了具有专业的财务管理、市场营销、经济法、管理学等基础知识之外，还要有一定的职业道德和素养。为此，在征信人才培训过程中，有必要注重专业知识和从业道德的双重培养，全面提高从业人员的综合素质。

由此可见，社会征信体系是一个庞大的相互联系的紧密体系，包含众多的子体系相互作用，需要众多的社会力量相互联动，其中既有外界提供的法律、环境等因素作为保障，又有征信体系内部专业理论知识、技能作为支撑。只有在各个子体系均发挥作用的情况下，才可能真正体现社会征信体系的服务和保障作用，为信用经济的深入发展提供必要的支持和相关的服务。

3)　征信体系各个子系统之间的协调与完善机制

征信体系各个子系统之间是一个相互影响、相互作用的有机整体，其中每一个体系都发挥着各自应有的作用，如果其中任何一个体系出现运行障碍和错误，势必影响其他体系的运作效率，带来不必要的麻烦。如图 7.9 所示为征信体系各子系统的协调和完善示意图。

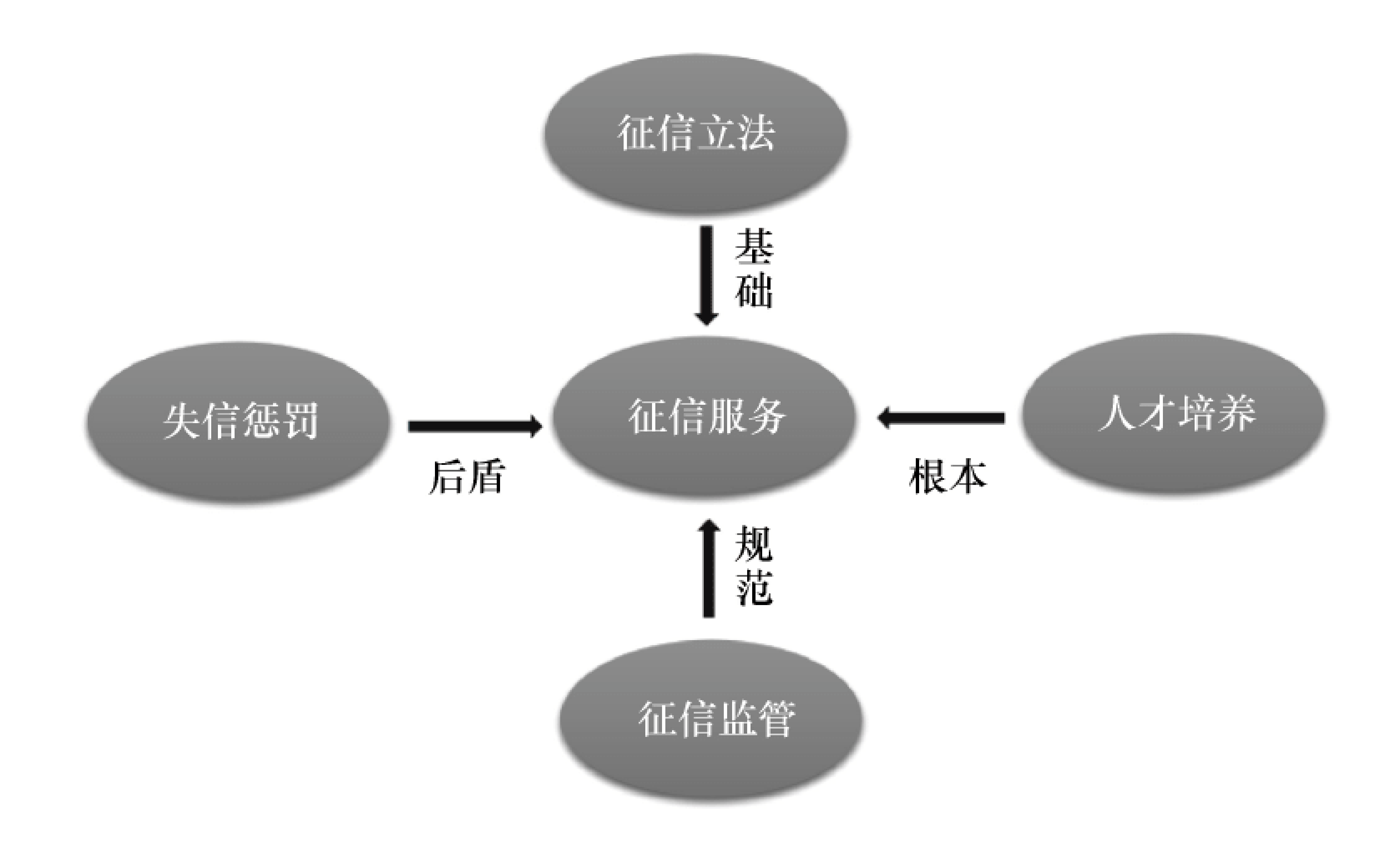


图 7.9　征信体系各子系统的协调和完善

(1) 征信立法系统是整个征信体系建设的基础和依据，只有完善的法律法规体系才能保障征信体系建设有章可循，有法可依。

许多国家的征信法律法规都明确地指出了征信数据采集的渠道和征信信息使用的方式，为征信体系建设从一开始就沿着良性轨道发展，奠定了必要的法律基础，同时也为征信监管有效的发挥作用提供了法律保障。一个国家的信用管理体系建设和征信服务的全面开展，首先必须创造必要的法制环境。要保障征信数据的开放，规范授信和信用管理行为，保护消费者的权益，就必须有一系列相关的法律法规及相应的惩罚机制。建立完善、高效的信用管理法制环境是信用行业健康规范发展的基础和必然要求，也是跨入征信国家的最主要标志。

(2) 征信监管体系对征信机构的征信行为进行必要的监督和管理，规范征信机构的行为，维护征信市场的秩序。

归纳起来，征信监管体系的主要功能有 6 个方面：第一，根据法律对不讲信用的责任人予以惩处；第二，教育民众在对失信责任人的惩罚期内，不要对其进行任何形式的授信；第三，在法定期限内，政府工商注册部门不允许有严重违约记录的企业法人和主要责任人注册新企业；第四，允许信用服务公司在法定的期限内，长期保存并传播失信人的原始不良记录；第五，对有违规行为的信用服务公司进行监督和处罚；第六，制定执行有关法案的具体规定。除此之外，征信行业的行业自律组织在各个机构自愿的前提下，依靠竞争规则和道德准则，使各成员自觉维护其权威性和制约性。与政府专设监管机构相比较，虽然监管效能和约束力度相对较差，但是在维护征信行业合理有序竞争和健康持续发展方面，同样具有自己特殊的效力和功能，也是征信监管必不可少的内容。

(3) 征信服务行业是征信体系的主力军，在征信法律和征信监管的框架下发展和壮大。

征信服务有两个特点，也是其两个优点：一是在被调查对象不知情的情况下完成，有助于保证调查结果的公正和客观；二是征信结果可以支持失信惩罚机制，对失信者给予必要的惩罚和制裁。其中主要的征信服务有企业资信调查(企业征信)、消费者信用调查(个人征信)、财产征信、市场调查和资信评估。征信体系的主体活动就是征信机构开展的各种征信服务，征信机构也是征信市场上最为活跃的市场主体，也正是征信机构之间的自由竞争和优胜劣汰，才保持着征信体系旺盛的生命力和无尽的活力。

(4) 失信惩罚机制是征信体系的坚强后盾，给失信者严厉的惩戒，给守信者实惠的奖励。

征信立法和征信监管侧重约束和规范征信机构的行为，保护被征信者的合法权益，但是征信的目的在于揭示信用信息，塑造良好的信用环境。所以，征信作用的发挥主要还是依靠激励机制和惩罚机制，只有奖惩分明，才能达到保护守信者，惩罚失信者的目的。因此，失信惩罚机制的存在，就像一道天然的防御屏障，对那些妨碍征信体系建设的失信者给予惩戒，同时激励守信者。

(5) 人才培养模式是征信体系建设的根本。

任何工作都需要有专业知识和技术的人才来完成，征信行业同样也不例外。征信体系建设归根结底要靠人才，离开了征信人才，再先进的法律、再有效的监管、再完善的服务也无法构筑适合中国目前现实国情的征信体系。当然,在注重征信专业技术知识教育的同时，千万不能忽视征信人员职业道德修养的培养，必须建设公正客观的征信人才队伍。

总之，征信体系各个子系统之间是相互作用、协调发展的整体，任何一个分支系统都对其他系统产生影响，在征信体系建设的过程中，关键不在于各个子系统分别建设得多么成功和完善，而在于处理好各个子系统之间的协调运作，相互配合。

下面以费埃哲(Fair Isaac Company)为例进行介绍。

FICO 信用分是由美国个人消费信用评估公司(Fair Isaac Company)开发出的一种个人信用评级法，已经得到社会广泛接受。

FICO 信用分是最常用的一种普通信用分。由于美国三大信用局(Experian、Equifax、TransUnion)都使用 FICO 信用分，每一份信用报告上都附有 FICO 信用分，以至于 FICO 信用分成为信用分的代名词。20 世纪 50 年代一位工程师 Bill Fair 和一位数学家 Earl Isaac 发明了一个信用分的统计模型，80 年代开始在美国流行。如今它是美国 Fair Isaac& Company 的专有产品，FICO 信用分由此得名。FICO 信用分模型利用高达 100 万的大样本数据，首先确定刻画消费者的信用、品德，以及支付能力的指标，再把各个指标分成若干个档次以及各个档次的得分，然后计算每个指标的加权，最后得到消费者的总得分。FICO 信用分的打分范围是 300～850 分。

虽然在审查各种信用贷款申请时，每个金融机构都有各自的方法和分数线，FICO 信用分可以帮助它们决策。然而信用分虽然可以作为发放贷款的决策工具，但不应当成为决策的唯一依据，更不能代替人的决策。

美国的法律禁止信用分作为拒绝消费贷款的唯一理由。一般地说，如果借款人的信用分达到 680 分以上，金融机构就可以认为借款人的信用卓著，可以毫不迟疑地同意发放贷款。如果借款人的信用分低于 620 分，金融机构或者要求借款人增加担保，或者干脆寻找各种理由拒绝贷款。如果借款人的信用分为 620～680 分，金融机构就要做进一步的调查核实，采用其他的信用分析工具，作个案处理。信用分低于 600 分，借款人违约的比例是 1/8；信用分为 700～800 分，违约率为 1/123；信用分高于 800 分，违约率为 1/1292。

FICO 信用分计算方法是把借款人过去的信用历史资料与数据库中的全体借款人的信用习惯相比较，检验借款人跟经常违约、随意透支、甚至申请破产等各种陷入财务困境的借款人的发展趋势是否相似。

如图 7.10 所示为 FICO 个人信用评分表。

FICO 评分模型中所关注的主要因素有 5 类，分别是客户的信用偿还历史、信用账户数、使用信用的年限、正在使用的信用类型、新开立的信用账户。评分权重占比如表 7.1 所示。

住房：自有 租赁 其他 无信息

25 15 10 17

现地址居住时(年)：＜0.5 0.5～2.49 2.5～6.49 6.5～10.49 ＞10.49 无信息

12 10 15 19 23 13

职务：专业人员 半专业 管理人员 办公室 蓝领 退休 其他 无信息

50 40 31 28 25 31 22 27

工龄：＜0.5 0.5～1.49 1.5～2.49 2.5～5.49 5.5～12.49 ＞12.5 退休 无信息

2 8 19 25 30 39 43 20

信用卡：无 非银行信用卡 主要贷记卡 两者都有 无回答 无信息

0 11 16 27 10 12

银行开户情况：个人支票 储蓄账户 两者都有 其他 无信息

5 10 20 11 9

债务收入比例：＜15% 15%～20% 26%～35% 36%～49% ＞50% 无信息

22 15 12 5 0 13

1年以内查询次数：0 1 2 3 4 5～9 无记录

3 11 3 −7 −7 −20 0

使用档案年限：＜0.5 1～2 3～4 5～7 ＞7

0 5 15 30 40

循环信用透支账户个数：0 1～2 3～5 ＞5

5 12 8 −4

信用额度利用率：0～15% 16%～30% 31%～40% 41%～50% ＞50%

15 5 −3 −10 −18

毁誉记录：无记录 有记录 轻微毁誉 第一满意线 第二满意线 第三满意线

0 −29 −14 17 24 29

图 7.10 FICO 个人信用评分表

表 7.1 FICO 评分规则

评分项	占比	评分规则
信用偿还历史	35%	信用账户的还款记录，包括信用卡、零售账户(从商户的赊购赊销情况获得)、分期偿还贷款、金融公司账户、抵押贷款；信用公开记录，主要包括破产记录、丧失抵押品赎回权记录、法律诉讼事件、留置权记录及判决；逾期偿还，包括逾期的天数、未偿还的金额、逾期还款的次数和逾期发生时距现在的时间等
信用账户数	30%	每个月仍需要偿还的信用账户数；仍需要偿还的分类账户数，如仍需要偿还的信用卡数量，分期付款账户数等；信用账户的余额；总信用额度的使用率，使用率越高，则说明客户的信用风险越大；分期付款账户偿还率
使用信用账户的年限	15%	一般来讲，使用信用账户的历史越长，越能增加 FICO 信用分，这项因素主要考虑平均信用账户账龄
新开立的信用账户	10%	客户的信用卡账户、零售账户、分期付款账户、金融公司账户和抵押贷款账户的混合使用情况，包括持有的信用账户类型和每种类型的信用账户数



续表

评分项	占比	评分规则
正在使用的信用类型	10%	新开立的信用账户类型及总数；新开立的信用账户账龄；信用查询申请数量，查询次数在信用报告中保存 2 年；贷款方查询客户信用距离现在的时间长短；最近的信用情况，对新开立的信用账户及时还款，会在一段时间后提高客户的 FICO 信用分

1) 偿还历史

影响 FICO 得分的最重要的因素是客户的信用偿还历史，大约占总影响因素的 35%。支付历史主要显示客户的历史偿还情况，以帮助贷款方了解该客户是否存在历史的逾期还款记录，主要包括以下几个方面。

(1) 各种信用账户的还款记录。包括信用卡(如 Visa、MasterCard、American Express、Discover)、零售账户(直接从商户获得的信用)、分期偿还贷款、金融公司账户、抵押贷款。

(2) 公开记录及支票存款记录。主要包括破产记录、丧失抵押品赎回权记录、法律诉讼事件、留置权记录及判决。涉及金额大的事件比金额小的对 FICO 得分的影响要大，同样的金额下，越晚发生的事件要比早发生的事件对得分的影响大。一般来讲，破产信息会在信用报告上记录 7～10 年。

(3) 逾期偿还的具体情况。包括逾期的天数、未偿还的金额、逾期还款的次数和逾期发生时距现在的时间长度等。

2) 信用账户数

该因素仅次于还款历史记录对得分的影响，占总影响因素的 30%。对贷款方来讲，一个客户有信用账户需要偿还贷款，并不意味着这个客户的信用风险高。相反，如果一个客户有限的还款能力被用尽，则说明这个客户存在很高的信用风险，有过度使用信用的可能，同时也就意味着他具有更高的逾期还款可能性。该类因素主要是分析对于一个客户，究竟多少个信用账户是足够多的，从而能够准确反映出客户的还款能力。

3) 使用信用的年限

该项因素占总影响因素的 15%。一般来讲，使用信用的历史越长，越能增加 FICO 信用得分。该项因素主要是指信用账户的账龄，既考虑最早开立的账户的账龄，也包括新开立的信用账户的账龄，以及平均信用账户账龄。据信用报告反映，美国最早开立的信用账户的平均账龄是 14 年，超过 25%的客户的信用历史长于 20 年，只有不足 5%的客户的信用历史小于 2 年。

4) 新开立的信用账户

该项因素占总影响因素的 10%。在现今的经济生活中，人们总是倾向于开立更多的信用账户，选择信用购物的消费方式，FICO 评分模型也将这种倾向体现在信用得分中。据调查，在很短时间内开立多个信用账户的客户具有更高的信用风险，尤其是那些信用历史不长的人。该项因素主要包括以下几个方面。

(1) 新开立的信用账户数，系统将记录客户新开立的账户类型及总数。

(2) 新开立的信用账户账龄。

(3) 目前的信用申请数量，该项内容主要由查询该客户信用的次数得出，查询次数在信用报告中只保存 2 年。

(4) 贷款方查询客户信用的时间长度。

(5) 最近的信用状况，对新开立的信用账户及时还款，会在一段时间后，提高客户的 FICO 得分。

5) 正在使用的信用类型

该项因素占总影响因素的 10%，主要分析客户的信用卡账户、零售账户、分期付款账户、金融公司账户和抵押贷款账户的混合使用情况，具体包括持有的信用账户类型和每种类型的信用账户数。

7.2 大数据征信

7.2.1 大数据征信概述

1. 大数据征信的含义

大数据征信是指运用大数据技术重新设计征信评价模型和算法，通过多维度的信用信息考察，形成对个人、企业、社会团体的信用评价。

大数据征信数据主要来源于网络上的公开数据、用户授权数据和第三方合作伙伴提供的数据。同时，互联网企业通过电商活动建立了宝贵的信用资源，从电商、微博等平台获取客户网络痕迹，从中判断借款人的信用等级，形成整体风险导向，完善大数据的积累。

大数据征信从其本质上来看是将大数据技术应用到征信活动中，突出强调的是处理数据的数量大、刻画信用的维度广、信用状况的动态呈现、交互性等特点，这些活动并未超出《征信业管理条例》中所界定的征信业务范围，本质上仍然是对信息的采集、整理、保存、加工和公布，只不过是以一种全新的方式、全新的视角来进行而已。

2. 大数据征信的特征与优势

互联网金融的业务一般都在线上完成，从申请到完成最快可能只需要几分钟的时间，而传统的征信流程时间长、进展效率低、业务覆盖面窄，已经无法满足越来越多的业务需求。大数据技术的发展，使信息来源收集到的一切可行数据都成为信用分析的基础，为互联网金融征信体系的建设指引了新的方向。

大数据征信相对于传统征信有以下几点特征与优势，如图 7.11 所示。

1) 依托互联网，覆盖范围大

关于收入情况、社保缴纳、信用卡消费等，与银行直接发生过借贷关系的人群，可以通过全国个人征信数据库查询到信用记录，从而进行相应的风险评估。但这一主要数据库牵涉面仍十分有限。在互联网上，只要个体有登记注册，开立银行账户，进行纳税，甚至社交等活动，便能用网络的痕迹，采取数据的深层挖掘与有效分析，同样也可能获得有价值的信用信息，这使征信人群辐射范围愈加扩大，得到延展。

传统征信		互联网大数据征信
收入情况，社保缴纳，信用卡消费等，与银行直接发生过借贷关系的人群	覆盖人群	只要个体有登记注册，开立银行账户，进行纳税，甚至社交等活动
主要使用传统结构化数据，其主要来源为借贷范畴	数据来源	除了现金流等财务数据外，根据互联网的活动痕迹，还可获知客户的交易行为、社会关系等半结构化的数据
离线的事后分析数据	时效性	在线实时更新
以财务数据为核心的小数据	数据量	数据海量化、维度化
企业运用场景较少，个人应用局限性大	应用场景	应用场景更加广泛，用户更加多元，需求多元

图 7.11　传统征信与大数据征信的区别

2)　获取广谱数据源，多方渗透

传统征信主要使用传统结构化数据，其主要来源为借贷范畴，而大数据征信不仅限于目前的形式，除了现金流等财务数据外，根据互联网的活动痕迹，还可获知客户的交易行为、社会关系等半结构化的数据。通过对这些半结构化数据甚至非结构化数据，进行不同维度、不同层次的挖掘与分析，可以得到关于人心理、行为、性格等根本的有价值的数据源，使之成为新数据的来源之一，继而纳入征信体系。由此可见，大数据提供的广泛而复杂的信息源对征信业务的信用评估渗透力与影响力十分强大。

3)　横向时间展开，实现数据实时性

离线的事后分析数据，让传统征信评价模式陷入了数据少、时效差的泥潭。在飞速的互联网+金融时代，只关注、分析考察对象历史信息早已不够。取代传统征信的精确性，大数据把重点转移至数据相关性方面。依靠大数据所具备的存量和热数据的典型特征，数据已成为一种在线实时更新的状态。在大数据征信的分析对象中，不仅包括考察目标的历史记录，还在时间的横向维度上加入当前信息。当数据的纵向挖掘与横向扩宽相结合时，信用评价的处理速度与决策效率将更加高效。

4) 多元变量，量化全面而精确

传统征信一般只针对以财务数据为核心的小数据构建单一变量。而科技的持续前进，让使用海量数据有了新的可能。大数据征信中信用评价模型可以容纳发展更多的变量，这为量化信用评价结果提供了全面而精确的保障，从而适应快速更迭的信息时代。

5) 人性化思路，适用多场景

传统征信体系的征信报告一般只有在信贷业务或者其他金融业务中用到，而大数据征信由于数据来源、内容模型思路主要来自借贷场景外的生活，如预订机票、酒店、租车等需要预授权支付或缴纳押金的场合，其得出的信用评价也更接近于人的本性的判断，基本人性化思路发展，有着可持续发展前景。

3. 大数据征信的难题

随着消费金融、网络借贷等互联网消费模式快速增长，以及大数据技术突飞猛进，大数据征信服务机构开始大量涌现。但多元化、多层次征信市场体系建设面临一系列挑战，有很多难题尚未破解。

1) 数据的质量、权威性问题

相比于央行征信系统的权威性、数据质量的高可靠性，大数据征信机构虽然数据来源更加宽泛、品种更加丰富，但数据质量、权威性受到质疑。美国国家消费者法律中心 2014 年 3 月对主要的大数据征信公司进行调查后发表了题为《大数据，个人信用评分的大失望》的调查报告，报告称，大数据征信公司的信息错误率高于 50%。这些公司的数据模型繁多又复杂，使用不准确的数据，有“垃圾进，垃圾出”之嫌。

2) 同人不同信用问题

决定大数据模型预测准确性的两个关键因素是数据和算法，各家征信机构的基因不同，数据来源不同。目前 8 家机构中，鹏远、中诚信、中智诚是传统型的征信机构，数据来源主要是金融数据、公共数据为主，而芝麻、腾讯、前海、考拉、华道则除了接入传统数据外，大量使用的是自身场景下积累的数据，这导致信用评估结果在不同公司间存在差异。

3) 个人隐私保护及信息安全问题

根据《征信业管理条例》规定，采集和应用个人征信信息必须要获得征信主体授权，商业银行在向人民银行征信中心报送和查询使用个人征信信息时，必须严格执行此规定，对于报送数据范围、查询用途范围、授权形式、异议处理等都有明确的界定。而大数据征信依赖大量个人的互联网交易记录、社交网络数据，在多重交易和多方接入的情况下，隐私保护的权利边界被淡化，隐私泄露风险被迅速放大，公民维护自己合法权益面临取证难、诉讼难等问题。

4) 公共信息的可获取、跨机构信息的可交换问题

如前分析，目前多家个人征信试点机构的信息来源带有浓厚的自身经营特点，申请个人征信试点机构大多首先拥有自己的具有垄断性的数据资源。而大数据征信要求的是信息的共享，而不是局部的垄断和壁垒。跨机构拥有的信息是否可交换，哪些需要获得信息主体的授权，如何保证交换过程和交换后信息不被滥用，在法律、监管、技术等方面都缺乏

标准。同时，工商、税务、司法等公共政务信息的可持续获取，尚得不到保证。目前的主要做法是，各家征信机构或信息使用机构分散地获取这类信息，获取成本高，数据质量和数据的可持续维护得不到保证。

5) 信息滥用带来的社会安全、公平交易问题

从首批试点的 8 家个人征信机构的运营情况看，市场开放之后，芝麻信用、腾讯征信、考拉征信等机构开始了一轮激烈的追逐赛，纷纷推出各自的评分产品，并争相在金融、购物、招聘、租车、租房、交友、酒店入住等领域尝试应用。但是，这些机构绘制出的人物“肖像”能否真实反映个人信用还令人质疑，获取信息所采用的关键技术的可靠性还有待进一步检验，没有制约的商业化应用很可能带来安全隐忧或消费歧视。

6) 征信机构的独立性问题

从各国征信机构的发展历程看，狭义的征信主要是为放贷机构的风险管理提供信息支持的活动，遵循“信息采集者与信息产生没有任何关系”的独立第三方原则。而目前试点的几家征信机构多不是独立的第三方：一方面，它们的数据来源于母公司，另一方面，其兄弟公司又涉足放贷业务。评分结果对于其各自经营领域的客户分析、风险判断具有强相关性，但其他应用场景下评分结果的相关性则有待验证。

7.2.2 大数据征信的理论基础

1. 大数据征信的经济学原理

1) 信息经济学理论

信息经济学是以“信息”为对象进行分析，优化资源配置，融经济学、管理学、运筹学、系统科学和信息科学于一体的交叉学科。信息经济学也是有关非对称信息下交易关系和契约安排的理论。交易双方是否诚实守信地履行契约约定的责任和义务反映着信息的不对称性，也决定了交易能否顺利进行，也决定了风险大小。

大数据征信的目的就是通过更多维度的信息分析总结为代理人提供更全面的参考，从而帮助代理人在合理的措施内，有效减少信息的不对称性，使风险降低。

信息不对称使得市场不透明，传统征信收集了银行系统内大量借贷数据，但覆盖人群不够，我国央行征信系统只有 3 亿多人有借贷历史，只占到中国 13 亿多人口的 20%多。美国也是如此，虽然三大征信局覆盖面较广，但还是有一部分人没有包含到。既然传统征信没有有效数据，就没法给那些不在其体系内的人进行信用评估，那么这些人需要借贷时就会从传统机构那里吃到闭门羹。而大数据征信是从互联网上用户的交易、社交等行为数据分析其信用资质。互联网时代用户在很多方面行为动作都自然而然用软件代替操作，势必留下了很多该个体的特征，利用数据模型分析出来以后，便能形成个体信用评价，某种程度上并不一定比传统征信的可靠性差多少。因此，大数据征信会使得信息对称度提高，信息经济学是大数据征信的核心理论之一。

2) 交易费用理论

交易费用理论核心在于节省交易费用，虽然企业和市场两种资源配置可以互相代替，但因为不确定性、小数目条件、机会主义及其存在有限理性有一定差异，致使交易费用节节高涨。交易费用的攀升会使得市场资源配置效率下降，所以尽量压缩交易成本对市场化

下组织结构和行为起着积极正面的作用。

大数据征信作为新型而有效的征信系统，从人力成本、高效率等众多方面大大节约了市场的交易成本。

传统征信因为某些原因给出较差信用评估时，往往给予较高的借贷费率。这就使得借款者的成本上升，这不利于经济合理发展。大数据征信有利于个体信息尽量对称，从而使得整个市场也趋向这种对称性，进而使得整体借贷费率趋于合理，这将促进经济按更真实情况发展。

3) 声誉理论

经济学中的声誉是指：在各方信息不对称的情况时，个体间存在一种信誉维持，这种维持会对双方起到一定的正面效用。存在相关合同时，交易行为可以经由法律加以限定，但在非正式合同的交易行为需要声誉来加以限制。较好信誉机制的形成有助于交易双方降低交易的成本，从长远看可以获得较好的利益。同时，授信方在良好信誉的关系中愿意为受信方提供更多信用服务，社会信用资源也能随之增加。

大数据征信的作用在于促使交易双方为了长远利益去维护声誉，从而形成稳定健康的信用大环境。

信用不佳会导致声誉下降，传统征信只在传统借贷范围内建立信用，但其实人们的声誉在其各个行为中都能表现出来。人们使用互联网的频繁度一定程度上已能反映其特征，声誉好坏也可以被分析出来，俗话说，人都是要面子的。大数据征信一定程度上也反映了人们在更多方面的声誉度如何，这会督促人们保持好声誉。

4) 长尾理论

长尾市场也称之为“利基市场”。“利基”一词是英文 Niche 的音译，意译为“壁龛”，有拾遗补阙或见缝插针的意思。菲利普 • 科特勒在《营销管理》中给利基下的定义为：利基是更窄地确定某些群体，这是一个小市场并且它的需要没有被服务好，或者说“有获取利益的基础”。

大数据征信市场的出现也是长尾理论创新应用之一，因为大量没有被服务到的小微群体数量非常庞大，而服务却没有跟上。

传统征信基本上对接大额借贷客户居多，对小额借贷不屑一顾，除了信息不全面问题，也有经济成本问题。而互联网的出现使得细分市场被挖掘，而大数据征信又更加针对分析这部分小微群体的行为痕迹特征。这部分群体数量非常大但单笔借款可能比较微小，但是乘积总和不可小觑。这就是长尾理论支持的海量不被传统机构重视的需求得以被挖掘和满足，而大数据征信正契合这点。

2. 大数据征信的管理学理论

1) 数据挖掘理论

在海量数据时代，征信系统需要利用数据挖掘技术对庞大数据进行提取分析，建立信用评分模型，从而运用到经济活动的各个环节中去。数据挖掘是一个交叉学科，涵盖了数学、统计学、机器学习、数据存储、AI 和高性能计算等多个学科。它需要有专业性人才参与发现大数据中有意义的模式与规律。

所以，建立健康的大数据征信体系的前提之一，是将其核心技术数据挖掘完善起来。

2) 信息加工理论

信息加工理论是因为问题解决和决策制定阶段时接收、理解、存储、使用信息的机制而成型的理论。当整个社会的企业与个人信息需要录入征信系统中时，就需要信息的加工。

整个征信过程是信息从接收到利用的过程，大数据征信对信息加工更加频繁。

传统征信的产生并不在互联网大数据等技术普及的时代。互联网和电子设备似乎已经成为人们的日常必备品，从而在这些基础设施上留下了人们的行为痕迹。现代科技的进步使得信息计算处理技术进入更高的层次，大数据征信离不开数据挖掘和信息加工，这都是对数据在技术上的处理高地。没有这些方面的支持，大数据也出现不了，自然也没有“大数据+征信”产生，所以这两者也是与经济学理论的信息经济学和交易费用理论一样，是大数据产生的关键原因。

3) 政府管制理论

为了维护、达到特定公共利益，政府可以出面进行管理和制约，这称为政府管制。政府管制的措施主要有审批、发放牌照、对企业限定经营范围等。有政府管制力量的介入可以维持一部分特定市场的行为。

由于征信机构牵涉众多个人或企业的利益，所以征信系统对政府管制的需求是双向的。一方面，整个征信体系需要政府的监督与管理；另一方面，征信行业也会谨防管制过度而阻碍整体的发展。作为征信系统的一部分，大数据征信对政府管制的应用与实践和传统征信亦一样。

大数据既然能采集挖掘人们各方面的特征，自然这些数据就有价值，价值带来两方面结果，一面是好，一面是坏，好的一面就是促进经济发展带来普惠，而坏的一面就会导致数据被滥用，隐私被泄露，使被征信主体可能遭受经济甚至其他损失。这些都与利益有关，那么就要进行法律法规制约，所以政府管制的角色作用就体现了，必须制定公平合理的监管措施。大数据和征信相关的行业法规的出台是必需的。

3. 大数据征信的社会科学理论

大数据征信不仅是涉及经济学或信息学的某一学科，同时还需要结合社会科学原理。心理学认为信用是指信任和安全感，是一种心理现象。伦理学中的信用是处理人际关系应当遵循的基本道德。

安全感和信任感某种程度上来自一方对另一方的信任，信任是非常重要的东西，信任也是了解对方特征以后做出的认可。按照之前所说，大数据征信能够一定程度地刻画出一个主体的特征。例如在借贷方面，假设分析结论是对方有意愿和能力还款，那么我们就应该予以信任把资金借给他。又例如在交朋友或婚恋方面，如果知道对方的信用度，产生的信任感会提升，因为认为对方是个靠谱的人，这也将一定程度改善人际关系，这点也可以反过来说，维护人际关系也需要提升自己的靠谱度。而大数据征信就是被量化的信用，满足社会对信任的需要。

所以总的来看，大数据征信综合了信息经济学、交易费用理论、声誉理论、长尾理论、数据挖掘理论、信息加工理论、政府管制理论以及社会科学理论。

7.2.3 大数据征信流程

征信大数据应用流程如图 7.12 所示。

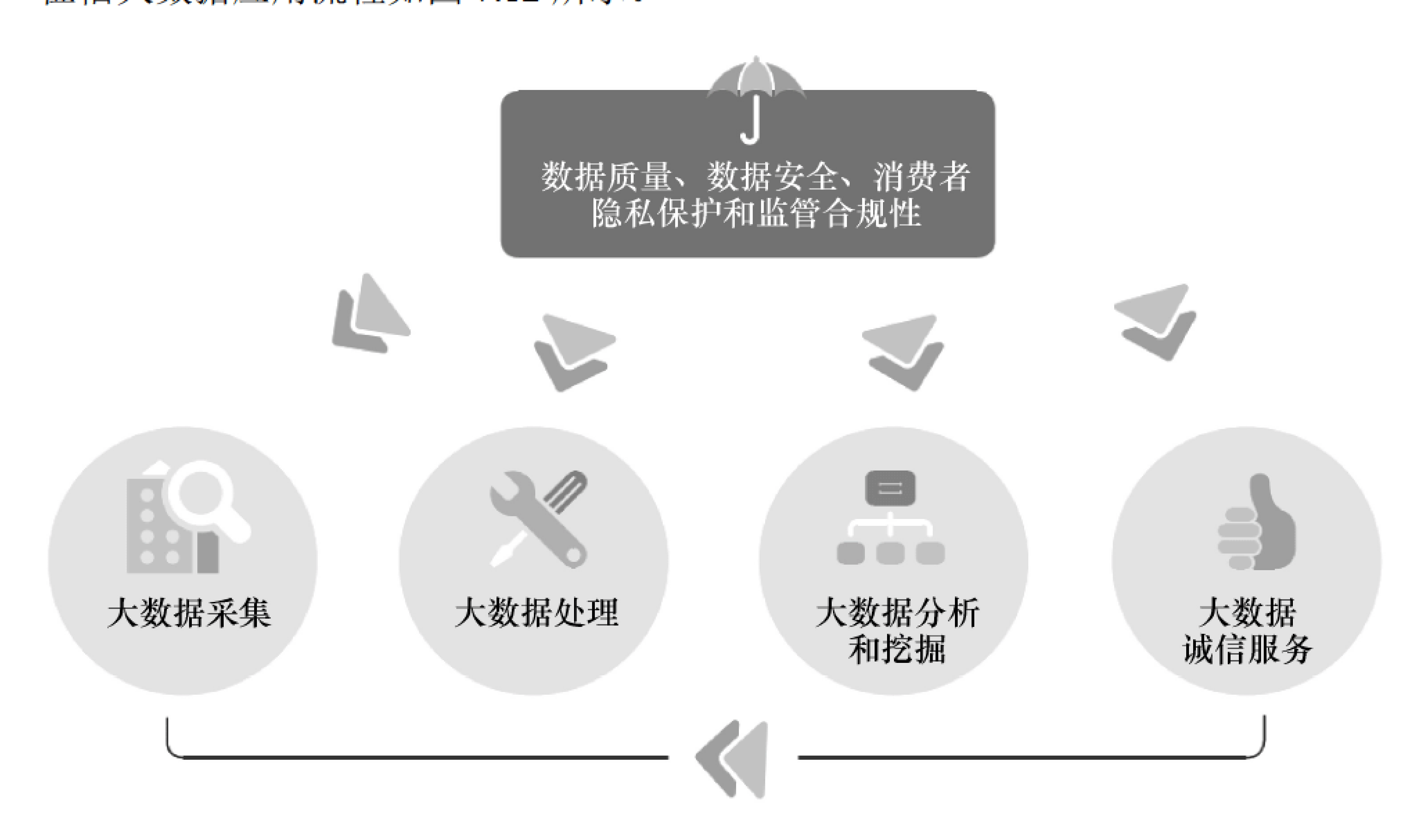


图 7.12 征信大数据应用流程

征信机构最基本的作用就是将分散在不同授信机构的碎片化的局部信息加工融合成为具有完整视图效果的全局信息，从中挖掘出风险信息，帮助解决交易过程中的信息不完整的问题，减少风险，降低交易成本，帮助商业机构更加有效地进行决策。大数据技术有助于对更加分散、碎片化、底层的数据加工处理成为更加完整的全局信息，更加有效地减少这种信息不对称。

类比于矿物加工提炼过程，征信机构的业务流程可以理解为将征信数据提炼为信用信息的过程，包括数据采集(数据可以理解为矿石原材料，数据采集可以理解为挖矿，收集矿石原材料)、数据处理(相当于矿石粗加工，去杂，粗加工成基本原材料)、数据分析和挖掘(矿石深加工，按照一定的配方，由不同生产线批量生产出不同的生活用品和化工用品)以及数据服务(对产品进行质量检查，进行包装，提供给各种终端用户)。

在大数据时代，大数据技术为征信发展提供了新的图景。大数据技术可以嵌套在整个征信的业务流程中，同时可以根据大数据服务的需求，不断更新和探索新的大数据来源。此外，征信大数据的处理流程中的每一个环节都要兼顾数据质量、数据安全、消费者隐私保护和监管合规性的要求。

7.3 大数据征信典型企业

7.3.1 国外大数据征信典型企业

如图 7.13 所示为美国征信体系。

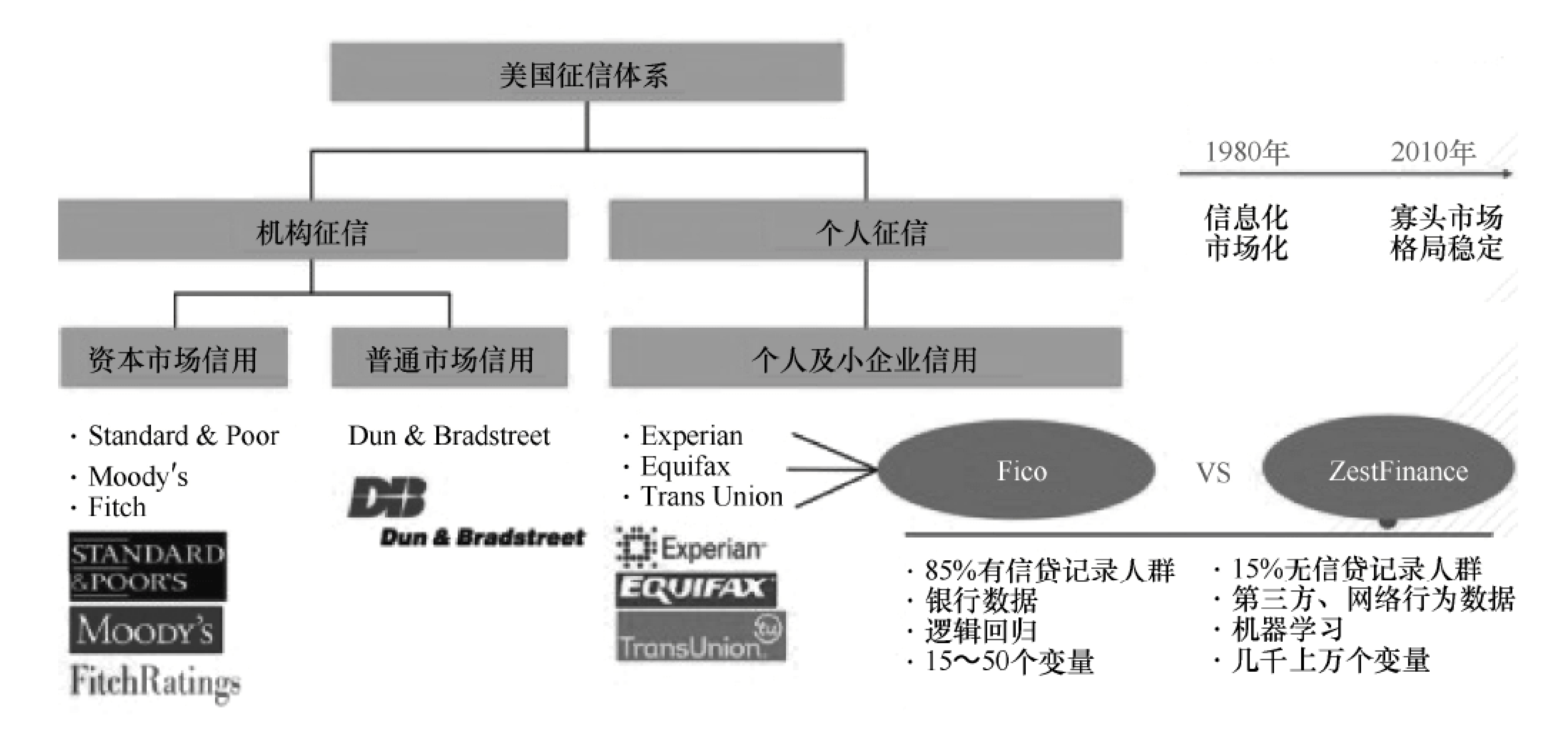


图 7.13　美国征信体系

美国在 1920 年前后征信体系即基本建立。目前，美国征信市场专业分工已非常清晰。整个征信体系分为机构征信和个人征信。

其中机构征信又分为资本市场信用和普通企业信用。资本市场信用机构包括 Moody’s、Standard & Poor’s、Fitch Ratings 等，普通企业信用机构包括 Dun & Bradstreet 等。

个人征信机构包括 Experian、Equifax、Trans Union 等。此外，美国征信体系中还有 400 多家区域性或专业性征信机构。

如图 7.14 所示为美国个人征信产业链。

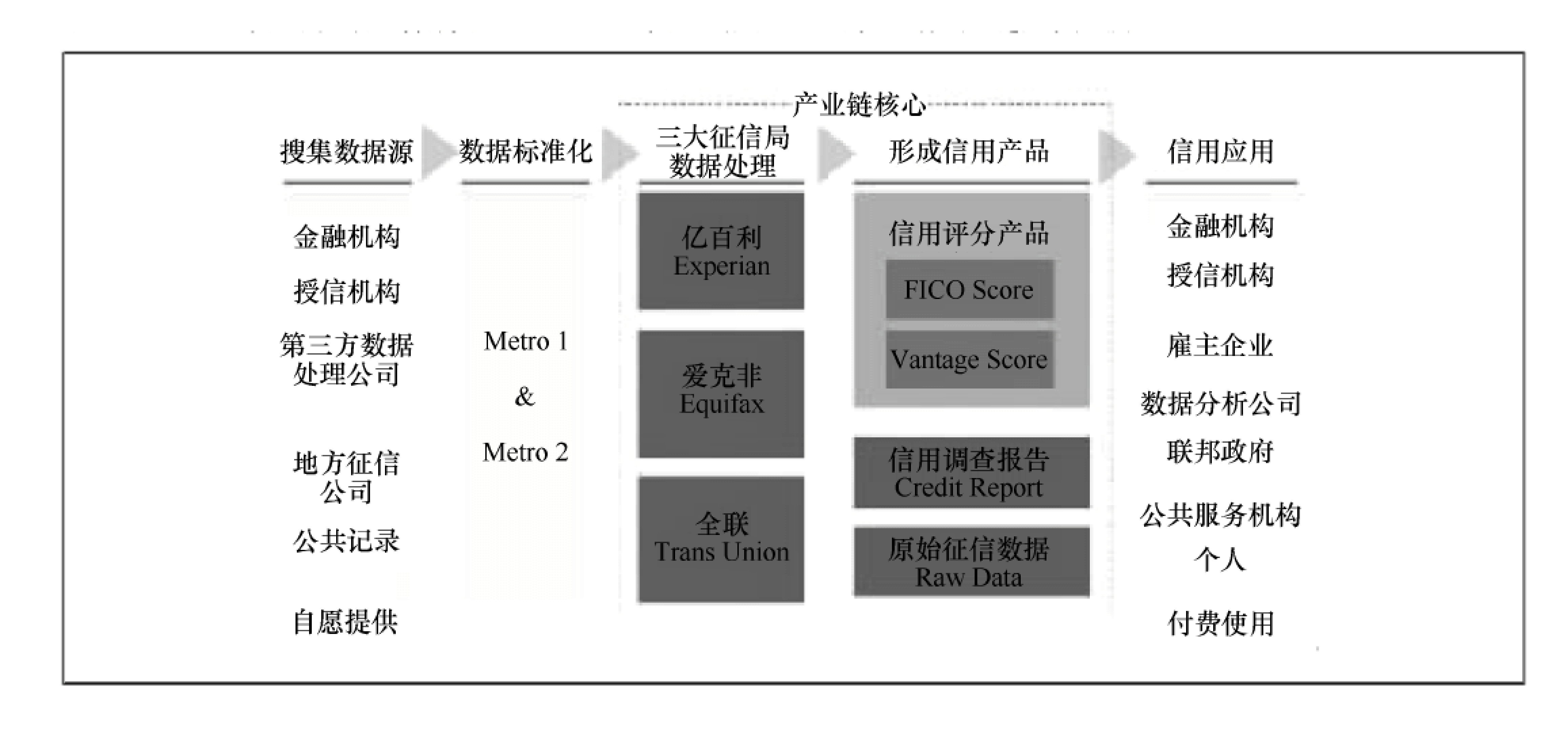


图 7.14　美国个人征信产业链

至今，美国个人征信体系已经形成成熟的产业链，其中三大征信局及相关信用产品是体系的核心，从 2000 年至今已形成三大征信局垄断局面。

1. 全联(Trans Union)

全联公司是美国的三大信用局之一，总部设在芝加哥，自 1988 年起开始提供美国全国性消费者信用调查报告。全联公司数据库中的 2.2 亿消费者资料，覆盖北美美国、加拿大、维尔京群岛和波多黎各。1990 年，全联公司已经拥有 45 家地区性信用局和 220 家代办处。

全联公司向全世界 50 多个国家提供 550 种产品和服务，信用报告的网上销售每年已达 4 亿次，其他传统方式查询更达 150 亿次。在数据的采集方面，全联公司拥有 7000 个数据供应机构，不间断地向它提供数据，从而使全联公司有能力、有资源每个月对 2.2 亿的客户资料进行 12 次数据更新，每次更新涉及 20 亿条的数据档案记录。庞大的数据加工系统，不仅是对计算机的硬件和软件的考验，也是对个人信用管理公司进行客户资料保密手段的考验。

1) 大数据采集：越来越多分散的、不同领域的数据源

大数据为征信活动提供了一个全新的视角，基于海量的、多样的、交叉互补的数据，征信机构可以获得信用主体及时、全方位的信息。

全联运营多年，已经建立了包含信息量丰富而且独特的数据源。全联目前拥有 30PB 数据，包括金融数据、信用数据、可替代数据(包括电信预付费、电商、社交网络、心理数据等)、身份数据、破产数据、抵押物数据、法院判决数据、保单数据、汽车数据以及从近 9000 个数据源抽取的其他数据，有全球超过 10 亿人的消费者客户的信息，从 2010 年开始以超过 25%的速度扩张。

全联的大数据资产，主要由以下 4 个方面组成。

(1) 传统的信用数据。

全联最基础的数据资产是信用数据库，基本上涵盖了所有美国信用活跃(有信用交易)人口的姓名、地址、现有信用关系和支付债务时间表。

该数据库中的信息是由成千上万的信用授予机构和数据提供商自愿提供的，传统的信用数据库支撑全联的基础的征信业务。

(2) 替代信用数据。

替代信用数据是指除了消费者信贷数据之外的信用交易相关数据，可以在消费者的信贷数据缺失的情况下，代替作为消费者信用描述和信用评估的手段，也可以作为一种增强信用评估的方法。

替代信用渠道(如租赁支付和公共设施支付)拓宽了传统信用数据库的范围。例如，全联拥有巴西最大的替代数据库，包括巴西联邦税务署的税务记录信息等上百个数据库和邮政编码(有 1.9 亿个人和 2900 万公司数据)。

(3) 消费者公共记录。

在国内央行个人征信系统，替代信用数据和消费者的公共记录统称为非银行征信数据。全联从法庭、政府机构和其他公共记录(如诉讼、抵押、判决、破产、专业许可、房地产、车辆所有权、其他资产、违规驾驶、犯罪记录和联络信息)中获取数据。

例如，在印度可以获得国家选举登记处(7.5 亿记录)、国家 ID 数据库(超过 5 亿记录)的信息。2013 年 12 月，全联收购了 TLO 公司的资产(该公司利用个人鉴定、欺诈保护和债

务找回的公共记录数据开发了数据产品)。2014 年 11 月，全联收购德国 DHI 公司(交通违纪和犯罪审判的数据提供商)。在南非，全联获得交通设备数据库(包括超过 1800 万车辆记录和特殊车辆识别码，是南非最全面的车辆数据库)。

(4) 专有数据库。

全联用复杂的算法生成自己专有的数据库，提炼并对数据进行标准化。这些数据是区别于其他竞争对手的，包括驾驶员违章记录、医疗资格信息、商业数据和房租交付信息等信息。这些数据库更准确地说并不是全联自己生成的，而是和其他机构合作获取的，比如房租交付信息是和美国一家房屋租赁公司合作而取得其数据的。

上述征信数据库受到监控，定期更新、复核。全联通过每月近 36 亿条记录的更新量来保持数据的鲜活度。全联在选择这些大数据时，起码要满足以下标准：①合规性，满足征信机构监管要求(包括信息安全和消费者个人隐私方面的规定)；②数据是活数据，可以及时更新；③数据是可以信赖的数据，从可靠的、具有公信力的数据源获得；④数据质量要保证数据能够正常使用。此外征信机构的数据或信息在被商业机构使用的时候常常要和本地的大数据做进一步的整合才能更好发挥作用。

2) 大数据处理：强大的匹配连接能力和下一代技术

征信数据规模变大，更新加快，类型复杂，需要有别于传统工具的新技术方法来完成数据处理和分析任务。

全联有以下大数据处理能力。

(1) 基础大数据技术。

全联自主研发了基础征信大数据技术，这为快速执行全联的应用和解决方案的更新提供了灵活性。

全联目前已经利用 Ab Initio(大数据处理软件平台技术)、Hadoop(开源分布系统的基础架构，适合处理超大量的数据)、Netezze(IBM 基于数据仓库的分析技术)和其他一些大数据分析和可视化技术来应对海量的数据(30PB)、分散的数据源(90 000 个数据源)和不同的数据格式(超过 4000 多种数据格式)。

全联的大数据技术可以处理、组织和分析跨越多个运行系统、数据库和文件类型的海量数据，同时处理快速变化的结构化和非结构化数据，加上每天数十亿的交易和数以兆计的数据交换。全联的大数据技术提供了高度的适应性，高效率和客户定制化，对于全联的解决方案，配合一些专业技术(如图形化开发和业务规则环境)，可以方便地和客户的工作流程整合起来。

(2) 增强的数据匹配连接能力。

大数据的商业价值实现关键技术之一就是匹配、连接和整合不同类型、不同来源的数据，其原理如下：首先找到多个数据源中信息对应的消费者，然后匹配消费者具体的信息项，将可能存在冗余的信息项进行合并或剪裁，得到消费者的全面、统一的视图。

全联的数据匹配技术能够整合多个数据源，连接多种信息，产生新的数据集，更好地评估风险和进行数据挖掘。

例如，全联 TLOxp 解决方案利用数据匹配能力(来自不同数据源)来确认和调查不同人之间、资产之间、位置之间和业务之间的关系，提供尽职调查、威胁评估、身份验证、欺

诈预防和检测的解决方案。在巴西，全联利用数据匹配技术连通巴西联邦税务署(税务记录信息)等上百个公共数据库和邮政编码。在印度，全联可以获得征信机构 CIBIL[Credit Information Bureau (India) Limited]的消费者风险信息，该信用数据库包括超过 2 亿的个人消费者和超过 1000 万的企业主体的信息。

(3) 下一代技术。

全联正在投入研发以大数据为特征的下一代征信技术，希望通过下一代技术的转型继续提供面向企业和消费者的服务，使得数据吞吐量增加，数据匹配能力提高，有较高的适应能力和较低的运营成本、更高的效率，保证更快的市场响应，可以实现使数据建档、数据清洗、数据入库的速率提高 10 倍，并由非 IT 人员自助完成，大幅度降低新产品的生产周期。

(4) 新技术探索。

近期，全联和南非一家高科技公司共同筹建南非国家声纹库，研发声纹识别技术进行消费者身份识别和反欺诈。据称，这种基于声纹技术的身份验证技术比传统基于知识(也称为“钱包外问题”)验证消费者身份方法效率高 80%。全联对生物识别的前沿探索目前还处于早期的研发阶段。

3) 大数据挖掘和分析：挖掘潜在信息和模式，释放大数据价值

征信机构早期的征信数据挖掘外包给费埃哲公司(FICO)，最成功的案例是 FICO 信用评分。随着数据分析技术的提高和普及，全联和其他几家征信机构开始建立自己的分析师队伍，开发自己的评分产品。但是由于历史传统的原因，征信机构还和 FICO 公司继续合作，向商业机构提供信用评分服务。具体来说，全联和 FICO 的合作只是在某些国家，如美国和加拿大。但在其他国家和地区，如中国香港、南非，全联提供的所有包括信用风险分数在内的产品都是由全联自主开发的。

理论上讲，信息更多可以提供更好的风险评估；但在实际操作中，随着平台的多样化、商业模式多元化的不断深入，商业实体之间关联性的加强，风险和商业机会的复杂性也在不断增加。大数据技术可以在消费者或信贷产品(组合)水平上进行风险测量和管理，使信用审批和定价更加精确。《经济学人》曾对大数据在金融风险方面应用做过调查，其中大数据在防范信用卡欺诈和减少违约率方面效果最好。全联利用大数据分析技术解决来自多个信息渠道、复杂海量的信息处理问题，提高风险模型的预测能力和稳定性，以及实时响应速度，帮助它的顾客在信用和风险管理中做出及时的决策。

为了充分释放征信大数据的价值，全联已经通过在技术、工具和人力资源方面的投入来研发复杂和灵活的分析和决策能力。

(1) 开发新的分析技术。

全联的分析师利用下一代技术和数据匹配能力实时读取来自不同数据源的数据并分析这些数据。一般来说，分析师配备有不同的建模和分析工具箱(例如可视化和机器学习)，目标能够在一天之内利用自服务的数据接口产生模型开发、模型验证和用于客户分析的数据样本。例如利用大数据分析工具，全联 Credit-Vision 解决方案中对一个新的贷款组合建模，只需要不到 1 天的时间，而传统工具和技术则需要开发 4～5 周。

(2) 分析团队。

在大数据时代、征信业发展涉及海量数据的存储、加工、处理、分析，需要大量的经济学、数学、计算机等各类型的高级综合型专业人才。全联拥有经验丰富的分析团队(一般都是高级专业人士或者是博士学位获得者)，拥有大量的行业经验并且对消费者信用数据有着深厚的知识储备。

(3) 研发分析工具。

数据分析工具是挖掘和分析征信数据的通用的基础软件组件。全联开发的分析工具包括基本预测模型和评分、消费者细分、业务标杆比较、欺诈建模、运营最优化等，能够满足特定的客户需求。

4) 大数据服务：丰富多元化的数据产品、个性化的服务

征信大数据使提供更多的信息服务、面向更多领域成为可能，大数据之间的交叉融合拓宽了征信产品和服务的广度和深度。全联通过提供综合的数据，先进的分析技术和决策能力的服务，帮助客户提高效率、管理风险、降低成本和增加收入。大数据使全联征信产品更加丰富、多元、及时和动态化，考虑不同客户群体的细分需求，提供更加个性化、客户体验更好的征信信息服务。大数据使全联的服务范围更广阔，从面向金融服务业转向在保险、汽车、医疗护理、电信、零售、出租审查、消费和法律执行等经济和社会领域帮助顾客做出关于信用和风险管理的及时决策。

基于特别的数据资源、分析和决策服务，全联近期研发的征信大数据产品和服务示例如下。

(1) 面向金融机构的征信产品 Credit-Vision。

不同于传统的个人信用报告只提供当月时点数据的服务，该产品基于 30 个月的时间序列数据，向金融机构客户提供个人消费者风险随时间变化的速度和严重程度，更精确地划分了风险。其和传统的信用分析产品的最重要区别在于它利用的不仅仅是当月的数据，而且是包括过去 30 个月的数据，因此对顾客信用各个方面的预测性更为准确。

(2) 面向保险公司的征信产品 Driver-Risk。

整合至少 3 年的司机驾驶的违规记录和其他大数据，高效地识别司机违规的可能性，从独特的视角来考察司机风险，降低保险公司的成本。

(3) 面向商业机构的市场营销产品 Ad-Surety。

基于全联自身的大数据，利用 O2O(互联网数据和数据库数据)匹配技术，帮助机构用户从包含 1 亿 3500 万美国消费者网络中识别潜在顾客，显示其个人信息并且测算效果，增加了找到目标顾客的可能性。

(4) 面向商业机构用户的决策分析产品 Decision-Edge。

这是一款软件即服务的产品，允许商业机构客户在和消费者交互情况下识别并验证消费者用户，对数据和预测模型的结果进行解释，根据机构客户定义的消费者标准帮助实现实时和自动化的决策。

全联的大数据技术的应用是一个综合性过程，是从数据采集、数据处理、数据分析与挖掘到服务的一个一体化的过程。随着业务的发展，今后全联的征信大数据增长主要从两个维度延伸：海外征信业务的发展增加消费者的数目；数据源的不断扩充并快速增加消费

者的信用描述。

目前全联的大数据是以结构化数据为主，基本不涉及社交网络、微博、论坛、互联网行为数据等非结构化数据，当然这一方面与美国的数据专业化运营和数据开放的大环境有关，另外一个重要的原因在于，世界本质上是结构化的，风险和商业信息首先主要隐含在结构化的数据中。因此，征信大数据的研发应首先解决好结构化大数据的处理和分析问题，挖掘出主要的风险和商业信息。虽然和国内流行大数据征信比较起来略显保守，但是由于其深厚的数据资产和征信技术的积累，全联对大数据技术的应用整体来看是一个自然的过程，根据数据信用相关性逐步扩张，目前已经开始研发以声纹为代表的生物识别等这些未来和征信相关的大数据。

虽然大数据技术给全球个人征信机构(如全联)带来了很多变化，如数据量的增大、数据类型的增多、处理技术的提升、分析能力的增强、服务范围扩大和征信产品的丰富，但是并没有给这些征信机构带来业务上颠覆性的改变，商业模式并没有发生变化，主要商业内容还是从基础信用信息服务、市场营销、决策分析到消费者的信用管理与反欺诈服务等。不过正如每一次数据技术的突破都会给征信机构带来更多的创新和颠覆，例如数据库技术和数据挖掘技术，未来的大数据技术不仅会延伸以全联为代表的全球个人征信机构的信用信息服务的广度和深度，而且未来有可能会带来一些商业模式上的变革。

2. ZestFinance

ZestFinance，原名 ZestCash，是美国一家新兴的互联网金融公司，2009 年 9 月成立于洛杉矶。ZestFinance 的研发团队主要由数学家和计算机科学家组成，前期的业务主要通过 ZestCash 平台提供放贷服务，后来专注于提供信用评估服务，旨在利用大数据技术重塑审贷过程，为难以获得传统金融服务的个人创造可用的信用，降低他们的借贷成本。

ZestFinance 起初是为传统的发薪日贷款(Payday Loans)提供在线替代的产品。发薪日贷款因借款人承诺在发薪日还款而得名。由于美国传统的信用风险评估体系无法覆盖全部的人群，大约 15%的人因没有信用评分而被银行排斥在外，无法获得基本的信贷需求。

除了解决传统信用评估体系无法解决的无信用评分借贷问题，ZestFinance 还主要面向传统信用评估解决不好的领域，将信用分数低而借贷成本高的人群视为服务对象，利用大数据技术降低他们的信贷成本。ZestFinance 目前也正在向信用风险管理的其他领域纵深扩展。2014 年 ZestFinance 宣布推出基于大数据分析的收债评分，旨在为汽车金融、学生贷款、医疗贷款提供一种新的评分系统。

ZestFinance 的基本理念是认为一切数据都是和信用有关的，在能够获取的数据中尽可能地挖掘信用信息。ZestFinance 对大数据技术的应用主要从大数据采集和大数据分析两个层面为缺乏信用记录的人挖掘出信用。

1) 大数据采集技术

ZestFinance 以大数据技术为基础采集多源数据，一方面继承了传统征信体系的决策变量，重视深度挖掘授信对象的信贷历史。另一方面，将能够影响用户信贷水平的其他因素也考虑在内，如社交网络信息、用户申请信息等，从而实现了深度和广度的高度融合。

ZestFinance 的数据来源十分丰富，依赖于结构化数据的同时也导入了大量的非结构化

数据。另外，它还包括大量的非传统数据，如借款人的房租缴纳记录、典当行记录、网络数据信息等，甚至将借款人填写表格时使用大小写的习惯、在线提交申请之前是否阅读文字说明等极边缘的信息作为信用评价的考量因素。类似地，非常规数据是客观世界的传感器，反映了借款人真实的状态，是客户真实的社会网络的映射。只有充分考察借款人借款行为背后的线索及线索间的关联性，才能提供深度、有效的数据分析服务，降低贷款违约率。

如图 7.15 所示，ZestFinance 的数据来源的多元化体现在以下几个方面。

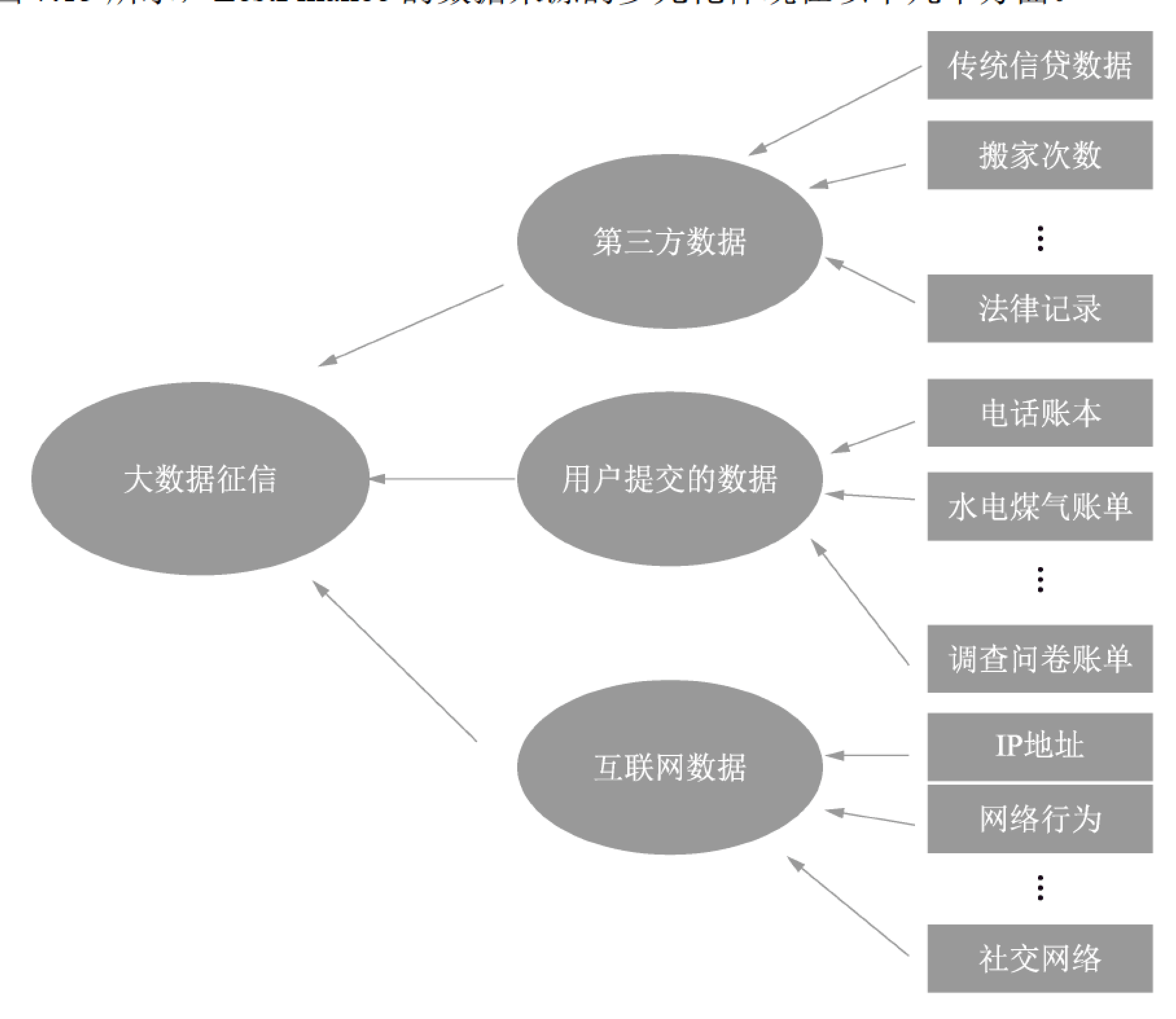


图 7.15　ZestFinance 的大数据源

(1)　对于 ZestFinance 进行信用评估最重要的数据还是通过购买或者交换来自第三方的数据，既包含银行和信用卡数据，也包括法律记录、搬家次数等非传统数据。

(2)　网络数据，如 IP 地址、浏览器版本甚至电脑的屏幕分辨率，这些数据可以挖掘出用户的位置信息、性格和行为特征，有利于评估信贷风险。此外社交网络数据也是大数据征信的重要数据源。

(3)　直接询问用户。为了证明自己的还款能力，用户会有详细、准确回答的激励，另外用户还会提交相关的公共记录的凭证，如水电气账单、手机账单等。

多维度的征信大数据可以使得 ZestFinance 能够不完全依赖于传统的征信体系，对个人消费者从不同的角度进行描述和进一步深入地量化信用评估。

2) 大数据分析模型

图 7.16 展示了 ZestFinance 的信用评估分析原理，融合多源信息，采用了先进机器学习的预测模型和集成学习的策略，进行大数据挖掘。

首先，数千种来源于第三方(如电话账单、租赁历史等)和借贷者的原始数据将被输入系统。其次，寻找数据间的关联性并对数据进行转换。再次，在关联性的基础上将变量重新整合成较大的测量指标，每一种变量反映借款人的某一方面特点，如诈骗概率、长期和短期内的信用风险和偿还能力等。然后将这些较大的变量输入到不同的数据分析模型中去。最后，将每一个模型输出的结论按照模型投票的原则，形成最终的信用分数。

■ 模型首先考虑成千上万种变量

■ 模型清理变量关系，转换变量为有用的格式

■ 被转换变量合并到元变量之中，描绘一个借贷者的某些特定方面

■ 元变量输入到不同的模块中，每一个模块代表一种“技能”

■ 每一个模块贡献一定分数比例，合成最终的分数

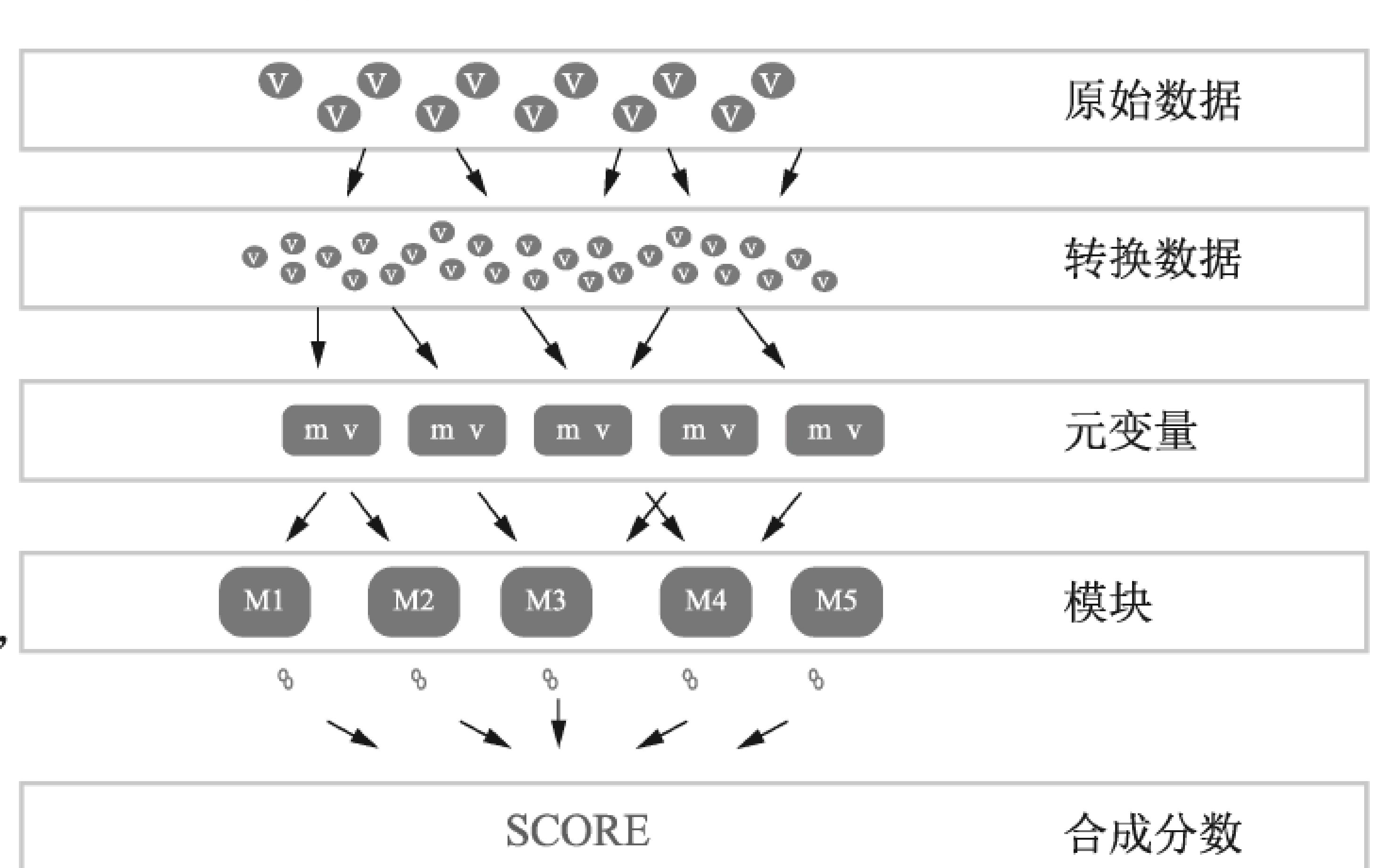


图 7.16 ZestFinance 的信用评估分析原理

其中，ZestFinance 开发了 10 个基于机器学习的分析模型，对每位信贷申请人的超过 1 万条数据信息进行分析，并得出超过 7 万个可对其行为做出测量的指标，在 5 秒钟内就能全部完成。这 10 个模型以如下的方式进行投票：让你最聪明的 10 个朋友坐在一张桌子旁，然后询问他们对某一件事情的意见。这种机制的决策性能远远好于业界的平均水平。

如表 7.2 所示，将这种基于大数据技术的信用评估体系和传统信用评估体系(以美国的征信体系为例)相比，发现主要的区别有以下几个方面。

表 7.2 传统的信用风险评估体系和基于大数据的信用评估体系的比较

	传统信用风险评估体系	基于大数据的信用风险评估体系
代表企业	FICO	ZestFinance
服务人数	有丰富信贷记录的人群(约占 85%)	缺乏或无信贷记录的人群(约占 15%)
数据格式	结构化数据	结构化数据+非结构化数据
数据类型	信贷数据	信贷数据、网络数据、社交数据
理论基础	逻辑回归	机器学习
变量特征	还款记录、金额、贷款类别	传统数据、邮箱姓名、填表习惯、浏览记录等网络行为

续表

	传统信用风险评估体系	基于大数据的信用风险评估体系
数据来源	银行提交给第三方的数据和银行系统内数据	第三方数据(如电话费账单、租赁历史等)和借款人自身提供的数据
变量个数	不到 50 条(变量库 400～1000)	多达上万条

(1) 服务的人群。新的信用评估体系可以服务没有被传统征信体系覆盖的人群，即没有征信记录的人群(美国的征信体系能够覆盖 85%的人群，覆盖不到 15%的人群)。

(2) 数据源。这种新的信用风险评估体系大量采用非传统的信用数据，包括互联网上的行为数据和关系数据，传统的信用数据(银行信贷数据)的比重仅占到了 40%，甚至完全不用传统的信贷信用数据进行风险评估。

(3) 关注的侧重点。传统的信用评估模型更关注授信对象的历史信息，致力于深度挖掘。而新的信用评估体系更看重用户现在的信息，致力于横向拓展。

(4) 信用量化评估的方式。新的信用评估体系抛弃了只用很少变量的 FICO 信用评分模型，基于大数据技术，不仅采用机器学习的模型，而且使用更多变量，一方面可以使信用评估的决策效率提高，另一方面还明显降低了风险违约率。

7.3.2 国内大数据征信典型企业

1. 芝麻信用：侧重电商

蚂蚁金服征信模式的运行机制是一个循环过程，自成体系。其运行过程如图 7.17 所示。

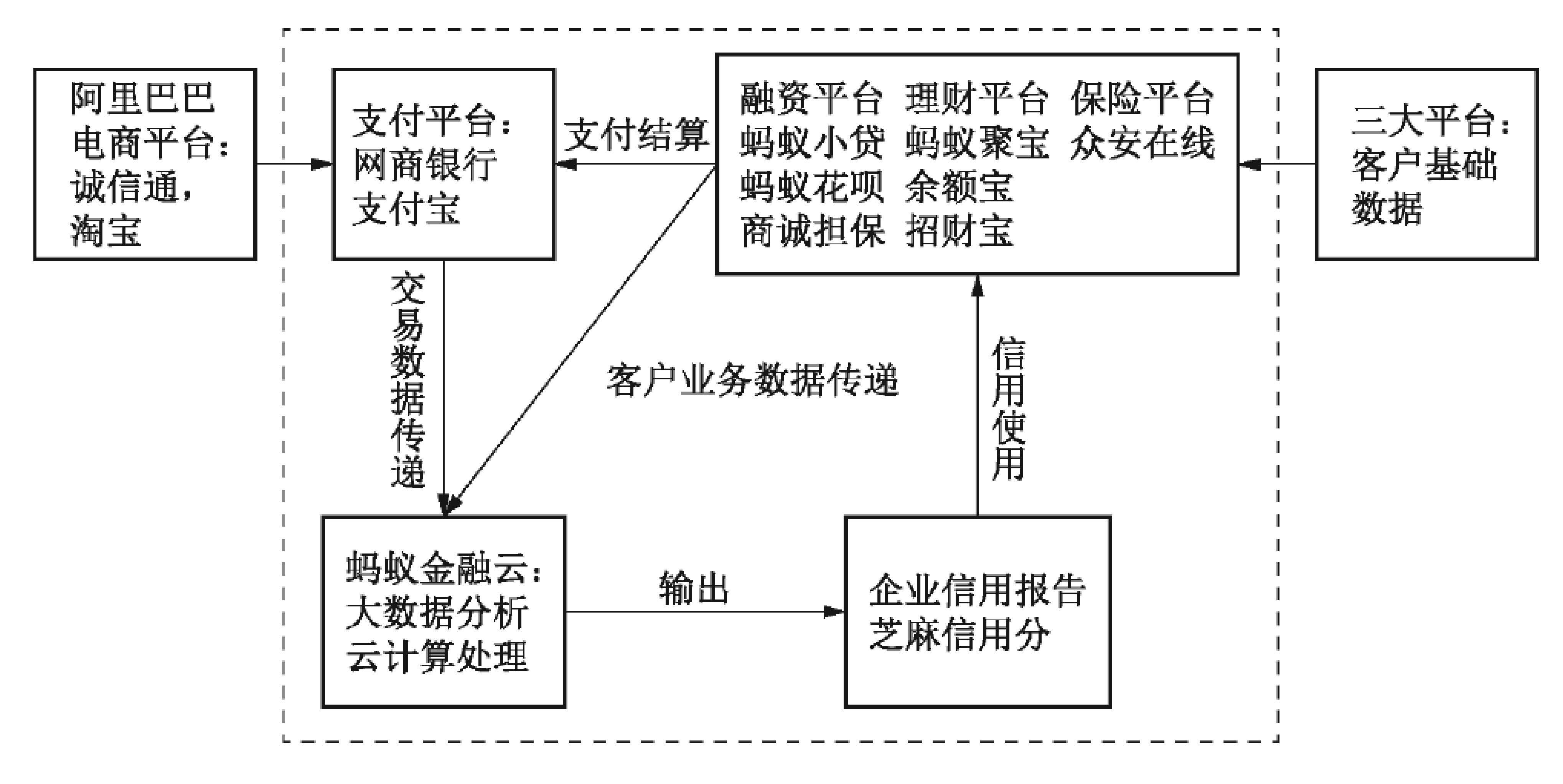


图 7.17 蚂蚁金服征信模式的运行机制

蚂蚁金服旗下拥有四大平台，即支付平台、融资平台、理财平台和保险平台。以阿里巴巴为依托，其诚信通和淘宝中个人和企业的交易数据会通过支付宝收录到支付平台，再将支付数据传递输出给蚂蚁金融云大数据库。融资、理财、保险三大平台以自身的客户数

据为基础，一方面将操作过程中的客户业务数据传递到蚂蚁金融云大数据库，另一方面也会通过支付平台来进行支付结算，而这部分交易数据也会随同支付平台输出到大数据库。蚂蚁金融云专注于云计算领域大数据的研究和研发，可以把各行为主体纷繁复杂的信息数据映射为其自身详细的信用评价，形成芝麻信用分和企业信用报告。

芝麻信用作为蚂蚁金服旗下独立的第三方征信机构，通过云计算、机器学习等技术客观呈现个人的信用状况，已经在信用卡、消费金融、融资租赁、酒店、租房、出行、婚恋、分类信息、学生服务、公共事业服务等上百个场景为用户、商户提供信用服务。

1) 大数据来源

其数据主要来源于以下 3 个方面。

(1) 阿里体系内的数据。包括阿里巴巴体系(淘宝、天猫)的电商交易数据和蚂蚁金服的金融数据。

(2) 外部合作机构提供的数据。主要有两种方式，政府方面的数据以购买方式获取为主，包括工商、学历学籍部门、法院、公安、电力、煤气公司等公共事业机构。另外，一些本身具有大数据积累的商业公司也是芝麻信用的合作对象，比如运营商、P2P 公司等，这部分通过合作、置换、服务输出等方式获得。

(3) 用户自主上传的信用数据。芝麻在 2015 年 7 月上线了上传功能，用户可以主动上传个人信息，包括学历学籍、单位邮箱、职业信息、车辆信息和公积金 5 个方面。

目前，芝麻信用带有购物、金融和社交 3 种不同维度的数据，其接入的外部数据源在八成以上，而阿里的数据源已减少至不足两成。

2) 大数据处理技术

芝麻信用在构建信用评分模型体系时，利用云计算、机器学习等技术，能以较低的成本对海量数据的关联性进行分析，还在充分研究和吸收传统征信评分模型算法的优势的基础上，积极尝试前沿的随机森林、决策树、神经网络等模型算法，挖掘出和信用表现有稳定关联的特征，从而更加高效和科学地发现大数据中蕴含的信用评估价值。

目前，芝麻信用应用了一种改进的树模型 GBDT，深入挖掘特征之间的关联性，衍生出具备较强信用预测能力的组合特征，并将该组合特征与原始特征一起使用逻辑回归线性算法进行训练，从而获得一个具备可解释性的准确的线性预测模型。

3) 大数据产品与服务

芝麻信用体系包括芝麻信用评分、信用报告、反欺诈、行业关注名单等一系列信用产品，提供反欺诈 IVS 信息验证服务(基于实名用户的欺诈风险识别，帮助提升合作伙伴反欺诈识别能力)、芝麻数据变量服务 DAS(还原用户画像，个性化的策略模型)、负面信息披露、还款提醒等服务。

芝麻信用评分即芝麻分是芝麻信用产品中的核心产品，并为用户提供信用评分服务。芝麻分一个看似简单的分数，背后是芝麻信用对海量信息数据的综合处理和评估。

2015 年 1 月芝麻信用开始在部分用户中进行公测，并推出芝麻信用分，这是我国首个个人信用评分。

芝麻信用分与国际通行的信用评分类似，分区间设定为 350 分至 950 分，分数越高代表信用程度越好，违约可能性越低。芝麻信用与 FICO 评分区别如表 7.3 所示。

表 7.3　芝麻信用分与 FICO 评分的比较

	芝麻信用分	FICO 评分
评分区间	350～950 分	300～850 分
评分维度	5 个，包括信用历史、行为偏好、履约能力、身份特质、人脉关系	5 个，包括信用偿还历史、信用账户数、信用使用年限、正在使用的信用类型、新开立的信用账户
评分等级	由低到高划分为 5 级：极差(350～550)、中等(550～600)、良好(600～650)、优秀(650～700)、极好(700～950)	不具体划分等级，一般而言，680 分以上代表信用状况卓著，620 分以下代表信用状况极差，620～680 之间，信用状况还需要做进一步核查
应用领域	目前在与芝麻信用开展合作的商户以及部分个人消费金融领域中应用	评分结果被美国三大个人征信机构采用，广泛应用于金融、通信、公共服务、日常生活等领域

芝麻分综合考虑了个人用户的信用历史、行为偏好、履约能力、身份特质、人脉关系 5 个维度的信息。

(1) 信用历史。过往信用账户还款记录及信用账户历史。目前这一块内容大多来自支付宝，特别是支付宝转账和用支付宝还信用卡的历史。

(2) 行为偏好。在购物、缴费、转账、理财等活动中的偏好及稳定性。比如一个人每天打游戏 10 小时，那么就会被认为是无所事事；如果一个人经常买纸尿裤，那这个人便被认为已为人父母，相对更有责任心。

(3) 履约能力。包括享用各类信用服务并确保及时履约，例如租车是否按时归还，水电煤气是否按时交费等。

(4) 身份特质。在使用相关服务过程中留下的足够丰富和可靠的个人基本信息。包括从公安、学历学籍、工商、法院等公共部门获得的个人资料，未来甚至可能包括根据开车习惯、敲击键盘速度等推测出的个人性格。

(5) 人脉关系。好友的身份特征以及跟好友互动的程度。根据“物以类聚人以群分”的理论，通过转账关系、校友关系等作为评判个人信用的依据之一。其采用的人脉关系、性格特征等新型变量能否客观反映个人信用，但目前还没有将社交聊天内容、点赞等纳入参考。

2. 腾讯征信：侧重电商

腾讯征信是首批经人民银行批准开展征信业务的机构之一，专注于身份识别、反欺诈、信用评估服务，帮助企业控制风险、远离欺诈、挖掘客户，切实推动普惠金融。

1) 大数据来源

腾讯征信依托于腾讯集团，信用信息主要来自社交、游戏、电商及第三方支付平台和合作平台。其中主要运用社交网络上海量信息，比如在线、财产、消费、社交等情况，为用户建立基于互联网信息的征信报告。腾讯庞大的客户群体为腾讯征信提供了海量信息。

2) 大数据处理与分析

腾讯征信由腾讯旗下财付通团队负责，通过其大数据平台 TDBANK，在不同数据源中，采集并处理包括即时通信、SNS、电商交易、虚拟消费、关系链、游戏行为、媒体行为、基础画像等数据，运用统计学、传统机器学习等方法综合考察用户的消费偏好、资产构成、身份属性和信用历史 4 个维度，得出用户信用评分，为用户建立基于互联网信息的个人征信报告。

3) 大数据服务

腾讯征信业务服务的对象主要包括两块：一是金融机构，通过提供互联网征信服务来帮助它们降低风险，能够为更多用户提供金融服务；二是服务普通用户，用很便捷的方式帮他们建立信用记录，这些信用记录能反过来帮助他们获得更多的金融服务。

4) 大数据产品

腾讯征信的征信产品主要分为两大类别：一类是反欺诈产品；另一类是信用评级产品。其中，反欺诈产品包括人脸识别和欺诈评测两个主要的应用场景。

(1) 人脸识别产品。

腾讯财付通与中国公安部所属的全国公民身份证号码查询服务中心，达成人像比对服务的战略合作。公民身份证查询中心，拥有全国所有公民的户籍信息，拥有国内最权威的身份信息数据库。双方通过深度合作，结合腾讯独创的技术算法，大力提升人脸识别的准确率及商业应用可用性，联手帮助传统金融行业解决用户身份核实、反欺诈、远程开户等难题。

人脸识别系统主要包括几个部分：人脸图像采集及人脸检测、人脸特征提取，以及特征相似度匹配与识别。

人脸识别技术能够应用的关键核心在于以下 3 点。

① 图像识别核心技术。

2015 年 1 月，腾讯的人脸识别技术正式登场亮相。腾讯对人脸识别研究由来已久，其旗下的承担人脸识别技术研发的优图团队，2014 年就已经在世界权威人脸检测评测集 FDDB 上达到世界第一水平，人脸识别 LFW 数据集准确率超过了 99.5%。在实际业务产品社交网络图像上的准确率高达 99%，对于身份证照片准确率甚至超过了 99.9%。在应用方面，腾讯的图像识别核心技术能力已积累了独有的优势。

② 丰富权威的样本数据库。

有效的图像样本库包括各类生活照和证件照，这是提升人脸识别技术的必要基础。

经过数年准备，腾讯采集标注了海量生活照训练样本数据。目前拥有世界上最大的黄种人人脸模型训练样本库，非常适用于国内环境。与此对应的用户人脸识别技术上已经有非常深厚的储备：在人脸检测、五官定位、特征提取和特征对比等关键步骤上，都已积累了世界顶尖的数据模型和算法。

最重要的证件照是身份证照片。腾讯财付通与公民身份证查询中心的深度合作，大力提升人脸识别的准确率及商业应用可用性。与其他几家公司的人脸识别技术不同的是，腾讯推出的人脸识别技术产品最重要的环节之一就是系统将用户的视频照、身份证照片跟公民身份证查询中心的权威数据三者做交叉验证，通过先进的算法和技术进行匹配，杜绝假冒身份的情况出现。

③ 广泛灵活便捷的应用场景。

传统金融中，用户在申请银行贷款或证券开户时，均必须到实体门店上做身份信息核实，完成面签。如今，通过人脸识别技术，用户只需要打开手机摄像头，自拍一张照片，系统将会做一个活体检测，并进行一系列的验证、匹配和判定，最终会判断这个照片是否是用户本人操作，完成身份核实。

腾讯与微众银行正在对金融、证券等业务进行人脸识别的应用尝试，相信不久之后，人脸识别将会出现在更多的应用场景里。

(2) 反欺诈核查产品。

腾讯征信旗下对公业务产品——账户级反欺诈产品已经开始接入合作机构，此款产品是国内首个利用互联网数据鉴别欺诈客户的产品，主要服务对象是银行、P2P、小贷公司、保险等机构。能帮助企业识别用户身份，发现恶意或者疑似欺诈客户，避免资金损失，支持国家的普惠金融政策。

(3) 信用评分及信用报告产品。

腾讯信用评分及报告则来自腾讯社交大数据优势，全面覆盖腾讯生态圈 8 亿活跃用户，通过先进大数据分析技术，准确量化信用风险，有效提供预测准确、性能稳定的信用评分体系及评估报告。

对于个人用户不但可以查询个人信用报告，还可以提高和完善自身信用情况，形成良性循环；对于银行等商业机构，该信用评分体系可以与自有体系形成交叉比对，帮助机构更准确地对用户个人信用做出判别，挖掘更多价值用户。通过多家金融机构实用验证证明，腾讯信用评分体系预测效果适用于银行，且评分性能稳定。

腾讯信用评分主要以星级的方式展现，共 7 颗星，亮星颗数越多表明信用评级越高。

星级主要由以下 4 个维度构成。

① 消费。用户在微信、手机 QQ 支付以及消费偏好。

② 财富。在腾讯产品内各资产的构成、理财记录。

③ 安全。财付通账户是否实名认证和数字认证。

④ 守约。消费贷款、信用卡、房贷是否按时还等。

3. 考拉征信：针对小微

考拉征信是由拉卡拉联合多家知名机构共同打造，作为独立的第三方信用评估及信用管理机构，考拉征信已同时获得了央行颁发的企业征信牌照和开展个人征信业务资质。而拉卡拉在征信方面的“抢跑”远不限于牌照——考拉征信不仅拥有国内首个专注于大数据征信模型研究的专业实验室，还是国内首家征信产品被银行接入的征信机构。

1) 大数据来源

考拉征信有着独特的 DNA，拥有多维度的数据来源。借助大数据技术和互联网平台，考拉征信汇集拉卡拉 10 年积累的便民、电商和金融数据，以及亿级个人用户和数百万线下商户的日常经营数据。此外，蓝色光标、拓尔思、梅泰诺、旋极、51job 等股东提供的相关数据、政府对外公开发布的公共机构数据，以及合作伙伴提供的个人及商户交易数据也同样能够为考拉征信提供有力的支持。

2) 大数据产品与服务

考拉征信现已成功推出个人征信、职业征信、商户征信等征信平台，为用户提供考拉个人信用分、商户信用分等系列产品，并为互联网金融行业提供了一整套信用评估体系及信用服务。目前，考拉征信业务已涵盖金融、民生、购物、租车、租房、交友等领域，与近 200 家机构开展了合作。

其中，考拉商户信用分是国内首款针对小微金融信贷及小微商户领域推出的征信产品，针对性地解决小微商户贷款难题，反映真实、整合和实时的商户运营情况。依托详尽的市场调研和信用数据验证的“商户信用分”，考拉征信已联合光大银行推出了“信盈卡”，创新性地推出了以“考拉商户分”换取“信用额度”的金融模式。基于“商户信用分”，小微商户通过考拉征信 APP 一键申请即可快速获得信用额度，这一创新产品为急需资金支持的小微商户带来了便利，特别是对民生领域的小超市、小百货、零售领域的商户提供了实实在在的融资支持。

而考拉个人信用分则是对个人用户信息进行加工、整理和计算后得出的信用评分，采用国际通行的信用分直观表现信用水平高低。分数的范围在 300 分到 850 分之间，分数越高代表信用程度越好。此外，作为国内首个开创职业征信平台的征信机构，考拉征信深刻理解“职场雾霾”现状和人才管理痛点，运用大数据征信技术发掘分析，可为企业提供即时、客观、全面的职业征信服务，帮助企业全面规避人才管理风险，提高人力资源效率。

不管是机构还是他人，要查看考拉分，都必须获得用户本人的授权。信用评估是直接以分值的形式呈现，以保护个人的具体信用信息和隐私。

4. 闪银：基于微信

北京闪银奇异科技有限公司，成立于 2014 年 4 月，是中国第一家互联网信用评估公司。其开发的“Wecash 闪银”(产品于 2013 年年底上线)是国内最先进的大数据信用评估系统。

闪银是一款基于微信、用大数据方式进行信用征集，利用数据分析技术和机器学习算法，进行快速授信、快速完成个人小额贷款的产品。

1) 大数据来源

主要根据用户在社交媒体、SNS 社区(如微信、微博、人人网等)发布的信息，分析用户在互联网上的行为轨迹及历史信息，并结合用户自主提交的身份信息、资产信息、网银流水等资料，对没有资信数据和借贷记录的用户人群进行信用风险评估。

2) 大数据分析

社交分析的具体过程如下。

首先分析微博、人人、微信朋友圈的社交数据，聚合形成对个人背景信息、社交活跃度、社交密度、社会影响力的评判。通过分析诸如“关注的人”“粉丝”“发布内容常用词”等信息，Wecash 能大体判断出一个用户的职业范围以及社会影响力等因素。

再结合用户上传的资产信息和银行流水等交叉验证。用户添加其微信公众账号(bank_9f)后，可直接在微信上提交社交网络地址、拍照上传必要的身份信息、资产信息、网银流水等资料，Wecash 随后通过其评估模型对个人完成信用评级，从而对个人完成最快

15 分钟的快速授信过程。授信后，提款、还款的功能均可通过微信完成。

这一过程通常仅需要 20 分钟。

5. 51 信用卡：侧重信用卡

51 信用卡主要是基于用户信用卡电子账单历史分析、电商及社交关系强交叉验证。根据用户的信用卡数据、开放给平台的电商数据所对应的购买行为、手机运营商的通话情况、登记信息等取得多维信息的交叉验证，确定用户的风险等级以及是否贷款给该用户(见表 7.4)。

表 7.4　51 信用卡客户风险等级模型

风险等级	账单管理时间	账单表现	手机入网期限	运营商	淘宝	最高额度
1	＞18 个月	授信卡数大于 3 张，单卡最高授信额度大于 3 万，额度使用率小于 50%，还款比 100%，账单完整度 100%，近 6 个月内极少延滞，近 3 个月内利息极少	＞5 年	近 4 个月有效通话记录大于 500 次，通讯录无负面联系人，与运营商匹配度高，关键联系人齐全	常用收货姓名及电话号码与申请人预留号码一致	5 万
2	＞12 个月	银行授信卡数大于 2 张，单卡最高授信额度(国有大于 1 万或商业大于 3 万)，额度使用率少于 70%，还款比大于 70%，账单完整度大于 75%，近 6 个月内较少延滞，近 3 个月内利息较少	＞3 年	近 4 个月有效通话记录大于 500 次，通讯录无负面联系人，与运营商匹配度高，关键联系人齐全	使用收货姓名及电话号码与申请人预留号码一致	2.5 万
3	＞6 个月	银行授信卡数大于 2 张，单卡最高授信额度(国有大于 1 万或商业大于 3 万)，额度使用率少于 70%，还款比大于 70%，账单完整度大于 75%，近 6 个月内极少延滞，近 3 个月内利息极少	＞2 年	近 4 个月有效通话记录大于 300 次，通讯录负面联系人很少，与运营商匹配度较高，有关键联系人	非常用收货姓名及电话号码与申请人预留号码一致，姓名对应手机号码大于 1 个	2 万
4	＞3 个月	银行授信卡数大于 1 张，单卡最高授信额度(国有大于 0.5 万或商业大于 1 万)，额度使用率高，还款比大于 10%，账单完整度大于 50%，近 6 个月内中等延滞，近 3 个月内利息中等	＞1 年	近 4 个月有效通话记录大于 50 次，通讯录有负面联系人较少，与运营商匹配部分匿名，无关键联系人	非常用收货姓名及电话号码与申请人预留一致，姓名对应手机号码大于 1 个，多个收货地址	1.6 万

续表

风险等级	账单管理时间	账单表现	手机入网期限	运营商	淘宝	最高额度
5	<1 周	银行授信卡数大于 1 张，单卡最高授信额度(国有大于 0.5 万或商业大于 0.5 万)，额度使用率高或超额使用，账单完整度大于 50%，近 6 个月内较多延滞，近 3 个月内利息较高	> 3 个月	近 4 个月有效通话记录大于 10 次，通讯录负面联系人较多，与运营商匹配较多匿名，无关键联系人	非常用收货姓名及电话号码与申请人预留一致，姓名对应手机号码大于 1 个，多个收货地址	0.8 万

51 信用卡风险等级由 5 个维度构成。

(1) 账单管理时间。信用卡有效存续时间越长，用户风险越低。

(2) 账单表现。根据用户的授信卡数、授信额度，以及还款比和账单完整度判断用户的还款能力和诚信程度。

(3) 手机入网期限。手机入网期限越长，用户风险越低。

(4) 运营商。通过近 4 个月有效通话记录以及通讯录中是否存在负面联系人判断用户自身的可靠程度。

(5) 淘宝。主要看常用收货姓名及电话号码是否与申请人预留号码一致。

本章总结

- 征信是指征信机构作为信用交易双方之外的独立第三方，收集、整理、保存、加工个人、法人及其他组织的信用信息，以在一定程度上揭示信息主体的信用风险状况，协助授信人或投资人进行风险管理的一种信息服务活动。简言之，征信的本质就是为授信机构或投资人的决策提供信息参考，是授信人或投资人之间的一种信息分享机制。
- 征信有六个方面的作用，它们分别是防范信用风险，促进信贷市场发展、服务其他授信市场，提高履约水平、加强金融监管和宏观调控，维护金融稳定、服务其他政府部门，提升执法效率、有效揭示风险，为市场参与各方提供决策依据、提高社会信用意识、维护社会稳定。
- 征信机构是负责管理信用信息共享的机构，从事个人和(或)企业信用信息的采集、加工处理，并为用户提供信用报告和其他基于征信系统数据的增值产品。从全球实践来看，征信机构一般分为三类：个人征信机构、信贷登记系统和企业征信机构。
- 征信体系是指与征信活动有关的法律规章、组织机构、市场管理、文化建设、宣传教育等共同构成的一个体系。征信体系的主要功能是为信贷市场服务，但同时具有较强的外延性，还向商品交易市场和劳动力市场提供服务。在实践中，征信

体系的主要参与者有征信机构、金融机构、企业、个人以及政府。征信体系模式主要有三种：市场主导型模式、政府主导型模式和会员制模式。

- 大数据征信是指运用大数据技术重新设计征信评价模型和算法，通过多维度的信用信息考察，形成对个人、企业、社会团体的信用评价。大数据征信从其本质上来看是将大数据技术应用到征信活动中，突出强调的是处理数据的数量大、刻画信用的维度广、信用状况的动态呈现、交互性等特点，本质上仍然是对信息的采集、整理、保存、加工和公布。

本章作业

1. 简述传统征信的四大原则。
2. 传统征信是如何进行分类的？
3. 请从理论角度阐述传统征信的作用。
4. 请简要绘出传统征信流程图。
5. 简述目前世界上存在的 3 种征信体系模式。
6. 我国目前的征信体系模式是哪一种？并阐述从何处做此判断。
7. 简述我国征信体系的各个子系统之间是如何协调的。
8. 大数据征信相比于传统征信有哪些优势？
9. 简述大数据征信大致流程。
10. 你认为在本章典型案例中的这几家企业的核心竞争力分别在哪里？如果现在你要创建一个大数据征信企业，你有哪些想法？

第 8 章

大数据与中国金融信息安全

本章目标

- 理解掌握金融信息安全的内涵、特征及重要性
- 掌握大数据对金融信息安全带来的机遇与挑战
- 了解我国金融信息安全的现状及制约因素
- 熟悉美国金融信息安全保障机制
- 掌握我国金融信息安全体系的构建策略

本章简介

继云计算、物联网被发明和应用之后，大数据成为当前信息产业的又一大技术创新。金融行业的大数据技术创新在给人们带来机会和挑战的同时也对现有的金融信息安全保护手段提出了更高的要求。

本章首先从金融信息安全的含义及特征属性入手，进而阐述金融信息安全的重要性。其次，引入金融大数据，简要描述了大数据给金融信息安全带来的机遇及挑战。再次，从宏观角度讲述了我国金融信息安全的现状及制约因素。最后，通过对美国金融信息安全保障机制的阐述，明确了我国金融信息安全体系构建的策略。

8.1 金融信息安全的重要性

8.1.1 金融信息安全的含义

1. 信息安全

信息安全的范围非常广泛，从国家层面讲，信息安全关系到国家利益和安全；从组织机构层面看，信息安全关系到组织的信息资产和商业机密，关系到机构的正常运作和持续发展；就个人而言，信息安全是个人隐私保护和个人财产安全的客观要求。

表 8.1 列举了业界对信息安全的定义。国际标准化组织对信息安全的定义更具通用性。

表 8.1 信息安全的概念

来源		定义内容
国际标准化组织(ISO)		为数据处理系统建立而采取的技术和管理的安全保护。保护计算机硬件、软件、数据不因偶然或恶意的原因而受到破坏、更改、泄露
美国国家安全电信和信息系统安全委员会(NSTISSC)		对信息系统以及使用、存储和传输信息的硬件的保护，是所采取的相关政策、认识、培训和教育以及技术等必要手段。确保存储或传送中的数据不被他人有意或无意地窃取与破坏，包括：信息设施及环境安全，如建筑物与周遭环境的安全;数据安全，确保数据不会被非法入侵者读取或破坏；程序安全，重视软件开发过程的品质及维护；系统安全，维护计算机系统正常动作
欧盟信息安全评价标准组织(ITSEC)		在既定的密级条件下，网络与信息系统抵御意外事件或恶意行为的能力。这些事件和行为将威胁所存储或传输的数据以及经由这些网络和系统所提供服务的可用性、真实性、完整性和机密性
信息安全资深专家	戴宗坤院士	确保以电磁信号为主要形式，在计算机网络系统中进行获取、处理、存储、传输和利用的信息内容，在各个物理位置、逻辑区域、存储和传输介质中，处于动态和静态过程中的机密性、完整性、可用性、可审查性和不可抵赖性，与人、网络、环境有关的技术和管理规程的有机集合
	沈昌祥院士	保护信息和信息系统不被未经授权的访问、使用、泄露、修改和破坏，为信息和信息系统提供保密性、完整性、可用性、可控制性和不可否认性

从微观角度看，信息安全主要是指信息生产、加工、传播、采集、处理直至提取利用等信息传输与使用全过程中的信息资源安全。信息安全的核心是信息处理过程的安全、信息存储环境的安全以及信息传输和数据交换过程的安全这 3 个方面。

从宏观角度看，信息安全是国家的信息化产业能力，以及信息技术体系能够抵御外来威胁与侵害，强调的是全面信息化产生的信息安全问题：一方面，泛指信息技术和信息系统发展的安全；另一方面，特指国家重要信息化体系(如国家金融信息系统、国家通信信息系统、国防信息系统等)的安全。

随着网络技术的发展，信息安全的内涵已经发展为信息系统运行安全、数据信息安全和通信网络安全，包括物理环境安全，软件、硬件和网络系统安全，信息保密安全，组织和个人隐私安全，信息系统基础设施与国家信息安全。

2. 金融信息安全

金融信息安全是指利用信息通信技术或者金融数据信息，对金融领域实施的各类安全威胁和应对手段。金融信息安全可能成为国家间网络安全对抗的战场。金融信息安全主要内容包括数据安全、运行安全、软件安全和物理安全。

1) 数据安全

没有数据安全就没有信息安全，数据安全管理必须贯穿数据生命周期的全过程。数据安全有相互对立的两方面的含义：一是数据本身的安全，主要是指采用现代密码算法对数据进行主动保护，如数据保密、数据完整性、双向强身份认证等；二是数据防护的安全，主要是采用现代信息存储手段对数据进行主动防护，如通过磁盘阵列、数据备份、异地容灾等手段保证数据的安全，数据安全是一种主动的包含措施，数据本身的安全必须基于可靠的加密算法与安全体系，主要是有对称算法与公开密钥密码体系两种。

金融业务的数据要求绝对安全和保密。 用户基本信息、用户支付信息、资金信息、业务处理信息、数据交换信息等的丢失、泄露和篡改都会使金融业遭受不可估量的损失。在互联网这样一个开放式的环境中，如何确保数据输入和传输的完整性、安全性和可靠性，如何防止对数据的非法篡改，如何实现对数据非法操作的监控与制止是互联网金融业务系统需要重点解决的问题。

2014 年，全国知名票务服务公司携程旅行网被曝其支付日志存在漏洞，用户银行卡信息可被黑客任意读取。这一事件引发大量用户更换信用卡，给社会公众造成巨大的恐慌，也对相关机构的信誉和作为互联网金融主力军之一的互联网支付蒙上了阴影。目前，很多互联网金融平台整体安全技术水平跟其业务的风险不匹配，加密系统和传输系统安全性并不完善，缺乏专业、核心的防范黑客攻击技术，一旦网络传输系统和环境被攻破，或者加密算法被黑客所破解，黑客就会乘虚而入，就会导致用户信息泄露、恶意冒充投资人进行恶意提现、大型 DDoS 攻击和 CC 攻击，以及来自黑客的恶意勒索。

2) 运行安全

运行安全主要是指金融各个信息系统能够正常工作，用户能够正常访问，系统之间的数据交换、调用等能够正常运行，避免出现运行不稳定、系统被攻击等现象。

3) 软件安全

软件安全主要是指互联网金融系统软件以及各个主机、服务器、工作站等设备中运行的软件的安全，避免软件的一些意外崩溃等。

4) 物理安全

物理安全是指各种硬件的安全，尽可能地减少一些不可抗力因素的影响。

8.1.2 金融信息安全的属性特征

金融信息安全除了具备广义信息安全的通用定义和特性外，还具有一些关键属性。根

据 ISO/IEC 27001：2013 信息安全管理体系，结合金融业对信息安全的主要需求，金融信息安全具有 9 个关键属性(见表 8.2)。

表 8.2　金融信息安全的属性特征

金融信息安全关键属性	属性描述
保密性	保障信息仅为那些被授权使用的人所获取。保证信息不被非授权访问，即使非授权用户得到信息也无法知晓信息内容或无法利用信息资源
完整性	保证数据从产生、传输到接收全过程的一致性，防止数据被非法篡改。涉及信息使用、传输、存储物过程中不发生篡改、丢失和错误；信息处理方法正确，不会对原始信息造成破坏
可用性	保障授权使用人在需要时可以获取和使用信息。保证合法用户对信息和资源的使用，而不会被不正当地拒绝
真实性	对信息的来源进行判断，能对伪造来源的信息予以鉴别
不可抵赖性	也称作不可否认性，通过建立有效的控制机制，防止相关方否认其行为，这一属性在金融信息安全中极其重要
可控制性	对信息的传播及内容具有控制能力。授权机构对信息的内容及传播具有控制能力，可以控制授权范围内的信息流向及其方法
可追溯性	对出现的安全问题提供调查的依据和手段。在信息交换过程结束后，相关方不能抵赖曾经做出的行为，也不能否认曾经发送/接收的信息
可靠性	信息系统在限定条件和限定时间内完成规定动作，可靠性是信息系统建设和运行的基本要求，也是金融信息安全的重要目标
连续性	具备应对风险进行自动调整和快速反应的能力，以保证关键业务的连续运转。金融业信息安全的连续性主要包括高可用性(high availability)、连续性(continuous operation)和灾难恢复(disaster recovery)

总的来说，金融信息安全研究的领域和范畴与一般信息安全有较多相似性，但从行业应用来看，更加注重涉及保密性、完整性、可用性、真实性、可追溯性、可靠性保护和连续性等方面的技术和理论。

8.1.3　金融信息安全的重要性

1. 金融安全是国家安全中重要而根本的内容之一

从某种意义来说，国家间的竞争和博弈，本质上是经济实力的竞争，其中，没有金融安全的保障，就没有国家发展的基础，甚至危及国家最基本的稳定。我国的金融系统信息化建设起步较晚，大量借鉴了国外的金融信息化发展模式，部署了大量国外提供的网络设备和主机设备、操作系统、中间件系统以及金融核心业务系统，同时，也大量采购了各种保障金融业务的咨询、方案、运维等服务内容。金融行业掌握着国家的命脉，金融安全关系国家安全。

然而，随着经济相互依赖性增强、信息通信技术快速发展、金融领域的逐步开放，以及新型金融业务的推广等，国家金融安全面临着与以往不同的风险，需要高度重视。

网络空间安全与金融安全的关系网络空间安全与金融安全密切相关，新一代信息技术所具有的融合、智能、宽带、移动、泛在等基本特点，以及“智慧城市”的推进和发展，使网络空间和所有的传统空间领域越来越深入地融合，网络空间安全直接关系到所有传统领域的安全。从国家发展的不同领域来说，政务、金融、国防、科技、社会稳定等各个方面，都离不开网络空间安全保障。因此，网络空间安全目前成为最受各国重视的热门话题。

在这种环境下，金融安全的保障，离不开对网络空间安全保障的深入理解和结合。从微观来看，电子商务、网上支付、网上银行甚至传统的信用卡等业务，受到交易过程安全的影响，也受到“网络钓鱼”等在线身份窃取类攻击的威胁，已经有大量案例。从宏观来看，传统的金融风险管理手段，不能完全涵盖恶意利用各类自动交易机制与系统(例如证券市场的量化投资)存在的缺陷，在短时间内将给国家造成巨大的金融损失，或者引起民众恐慌进而引发社会动荡风险。

近年来，网络空间安全形势十分严峻。我国每年都发现千万级左右的 IP 地址被境外攻击者秘密控制，大量重要网站的数据被大规模窃取。随着银联等金融业务走向国际，网络钓鱼攻击等身份窃取攻击转向我国银行等网站，移动互联网的快速发展和深入应用，使智能终端成为重点攻击目标，恶意应用增长迅猛，直接威胁用户经济利益及金融新业务推广。此外网络拒绝服务攻击十分活跃，针对我国信息基础设施的严重攻击事件时有发生，新型攻击手段不断出现，重要用户部门却对其了解不多。

目前，网络空间中的威胁因素高度复杂。网络空间安全大致经历了 4 个阶段。第一阶段是“白开心”，攻击者主要是“脚本小子(SCRIPT KIDS)”或纯粹的技术黑客，行为“损人不利己”，形式主要是计算机病毒、蠕虫和拒绝服务攻击等；第二阶段是“淘黑金”，标志是“趋利”，攻击者为各类计算机犯罪分子，攻击目标主要是商业性网站和用户，攻击形式是木马、网络钓鱼、拒绝服务攻击等 ；第三阶段是“窃密者”，攻击者是互联网窃密行为，不仅窃取商业秘密，还包括军事秘密、经济情报、科技情报等，攻击形式是通过木马、僵尸网络等，并结合社会工程学手段 ；2010 年始进入第四阶段，即“大玩家”阶段，具有政治动机并具备资源和能力优势的攻击者出现，国家势力成为攻击发起方之一。网络中混杂不同攻击者带来的安全威胁。相较而言，趋利性攻击在数量上占主流，但国家势力发起的攻击，隐蔽性和破坏性十分突出。不同的攻击者和动机会导致攻击目标、方法、拥有的资源、破坏能力等都有很大差异。因此，思想及安全保障能力要做大调整。

网络空间安全目前进入国家间对抗阶段。除技术层面出现只有国家力量才可能完成的高度复杂的攻击事件外，在战略、外交、产业等层面明显表现出“冷战”时期的特点。国际层面对抗抬头，信任降低，合作受影响。对“大玩家”拥有的手段、具备的能力、掌握的资源(包括战略级的漏洞或后门)、潜在动机等，已有网络安全保障力量了解甚少。基于经验、能力和机制，来应对当前国家间网络安全对抗中可能产生的威胁，是目前的最大风险。

对金融领域亦如此。针对金融领域的攻击，除网络钓鱼、金融诈骗、非法转账及其他

以获利为目的的漏洞利用行为之外，国家势力或恐怖分子还可能发起其他类型、充分利用其特殊资源、以破坏金融体系或造成重大损失为终目的的攻击。

2. 金融信息安全是国家发展战略的重要基石

金融是现代经济的核心，金融信息系统是国家重要的关键信息基础设施。金融信息安全不仅关系国家经济社会安全，也关系着金融企业的持续发展。金融信息安全无疑是国家发展战略的重要基石。

人类进入 21 世纪以后，信息安全问题变得日益重要，目前已经上升到国家发展战略层面，很多发达国家视其为仅次于恐怖袭击的重大安全领域。

互联网的普及和信息技术的发展给金融行业带来了前所未有的机遇，金融系统得到了蓬勃发展。计算机正越来越多地参与到金融系统活动当中，成为其不可或缺的一部分。而电子信息化也成为现代金融发展必然的趋势。但与此同时，信息技术的参与也在一定程度上削弱了交易的可控性，使交易风险大大增加。任何一个精通计算机或网络的人都有机会对金融系统进行蓄意破坏，人为的干预和破坏都会对金融系统带来重大的影响。在金融系统中，运行的数据基本上都是以资金信息为主，由于其庞大的用户基数并随着时间的积累逐渐形成了海量的数据，这些数据的存储和保护给人们带来了巨大的挑战。金融信息往往涉及国家、集体或个人的利益，一旦有数据损坏或者非法数据访问，都将造成不可弥补的经济损失。因此，金融信息安全正成为一个具有挑战性的命题。

随着我国信息化的不断推进，国家对信息安全工作的重视程度日益增加。2012 年 7 月，国务院发布了《关于大力推进信息化发展和切实保障信息安全的若干意见》，这是国家信息化建设和信息安全工作的纲领性文件，对于今后我国信息化建设和信息安全工作具有重大的指导意义。

8.2 大数据给我国金融信息安全带来的机遇和挑战

任何事物的发展都具有两面性。大数据的快速发展在为金融信息安全带来发展机遇的同时，也带来了一些挑战。

8.2.1 大数据给金融信息安全带来的机遇

大数据实现了对传统数据信息结构的解构，与传统数据结构相比成为一个具有流动性、信息共享与连接的数据池。

通过这种灵活的大数据技术，人们可以在最大程度上利用人们以为无法有效利用的数据信息形式来实现对金融企业的高效运营，为金融业的发展也带来了更大的机遇。大数据信息技术的提高也使得数据信息安全工具和技术有所发展，让金融信息安全的监督更加精细、高效与及时。

1. 对大数据的挖掘和应用将创造更多的价值

在大数据时代，大数据的发展重点已经从数据的存储与传输发展到了数据的挖掘和应

用，这将引起金融企业发展的商业模式的变化，并且能为金融企业带来直接的利润，也可以通过积极的反馈来增强金融企业的竞争力。

2. 大数据的安全愈发重要，为金融信息安全带来了发展机遇

在大数据时代下，金融信息安全事件发生的次数逐年增多，金融信息安全事件所引发的数据泄露并由此带来的经济损失也越来越大。

随着科学技术网络的不断进步，大数据安全不仅是金融企业需要面临和维护的对象，也是个人消费者要面对的对象。大数据已然渗透到我们生活的方方面面，这一切使得金融信息安全越来越重要。

大数据提高了金融数据信息的价值，但是数据信息安全意识薄弱以及频发的金融信息安全事件，对信息安全技术和工具均提出了更高的要求。目前所使用的信息安全技术、工具、管理手段以及相关的不能解决这个问题的方法、方式都应该得到改进，而大数据的发展为这一发展提供了巨大的可能性。所有这些，都为金融信息安全的发展提供了新的发展机遇。

3. 在大数据时代下，加快了信息安全的发展速度，云技术拥有巨大潜力

在大数据这条巨大的产业链中，参与者众多，面积也十分广泛。如果按照产品的基本形态来进行划分，可分为硬件、应用软件和基础软件三大类。云技术和金融信息安全纵贯这三大领域。纵观各个领域的国内外的发展情况，信息安全和商业智能的发展速度最快，尤其是云技术，它将有更大的发展潜能。这三者将成为大数据产业链的三大主要推动力。

8.2.2 大数据给我国金融信息安全带来的挑战

由于大数据参与金融业发展起步较晚，目前还不成熟。大数据金融并不都意味着机遇或者商业上的无限潜力，在我们能够很好地了解大数据、管理大数据之前，实际上还同时意味着巨大的风险。

1. 数据应用侵犯客户个人隐私

大数据技术的应用和隐私保护的价值的争议由来已久。目前，随着技术的高速发展，信息传递技术与超强的计算机系统使得数据高速分析成为可能。交叉检验技术和“块数据”技术的广泛应用，使得基于大数据的身份识别日益简单且难以察觉。近年来，大数据金融需要对客户信息进行全方位的分析与应用。但是，这些应用也容易跨越雷池，挖掘过多的私人信息，造成对客户隐私的侵犯。

2. 数据监听威胁国家金融安全

2013 年“棱镜门”事件表明，“海量数据+数据挖掘”的大数据监听模式可以对他国重要机构进行精确监听。无论是软硬件设施还是数据服务，我国金融企业都过度依赖国外厂商。在信息传输的各个环节，中国金融企业和金融机构的内部信息可能通过国外厂商预留的“后门”泄露给国外机构，从而成为大数据监听的受害者。

3. 虚假数据导致金融市场异常敏感

由于信用信息是互联网金融的纽带，是驱动业务的核心因素，因此，基于信用信息数据的金融决策对信息非常敏感，从而导致金融市场敏感。如果数据不准确，就可能导致错误的交易行为，并进一步引发金融市场风险。2013 年 4 月 23 日，美联社 Twitter 账号出现“白宫遭袭”的假新闻。受此影响，众多基金公司的交易程序自动抛售股票，美国股市随即暴跌。

4. 法律监管缺失存在风险

由于中国大数据金融发展时间较短，金融市场内现有的证券法、银行法、保险法等都是在传统金融模式的运营下制定的，面对大数据金融相关的金融创新产品，约束力不强，不能有效地适用于这一新生事物的需要，对大数据征信数据处理的各环节及个人隐私等问题未定义明确界限。

目前，我国金融信息安全的法律风险主要包括两个方面：一是金融信息安全法律法规不够健全；二是金融信息安全立法相对滞后和模糊。近年来，我国相继出台了《电子签名法》《网上银行业务管理暂行办法》《网上证券委托管理暂行办法》《证券账户非现场开户实施暂行办法》等法律法规，但这些法律法规也只是基于传统金融业务的网上服务制定的，并不能满足互联网金融发展的需求。因此，在利用互联网提供或接受金融服务时，配套法规的缺乏容易导致交易主体间的权利、义务不明确，增加相关交易行为及其结果的不确定性，导致交易风险增加，不利于互联网金融的健康发展。

5. 层出不穷的互联技术应用是当前金融信息安全面临的最大挑战

移动互联、云计算、下一代互联网、大数据等新兴技术的蓬勃发展，是催生互联网金融时代快速到来的主要推手。一方面，这些基于开放性网络的互联网金融服务，使得以往金融信息安全技术防范已经不能全部适应新互联网技术的进步速度；另一方面，这些新兴互联网技术自身还在不断发展，其技术成熟度还不稳定，特别是第三方支付、P2P 等互联网金融新业态还处于起步阶段，其信息安全管理水平不高。如何尽快建立一套既符合金融行业特点，又能快速跟进互联网新技术发展需要的金融信息安全技术规范显得十分紧迫。

6. 网络安全防控是互联网金融信息安全防范的难点

《2013 年中国互联网发展报告》中指出 2013 年互联网遭到的网络攻击同比增长 14%，已经连续多年呈上升趋势，其中涉及客户信用卡信息、各种资金账户信息的非法网络攻击行为增速位居前列。曾有专家说过，“互联网金融第一要素就是互联网，安全就是生命线”。由于互联网模糊了传统金融领域的界限，使得金融行为范畴借助互联网技术衍生到前所未有的新领域。一方面，无论是传统金融机构还是新生的互联网金融公司，来自互联网的各种入侵破坏行为已经成为日常信息安全防范的重点；另一方面，在互联网开放性的影响下，各类基于互联网平台的金融创新业务也带来一些类似网络洗钱和网上支付诈骗的社会安全问题，这类网络安全防控不断突破传统金融安全的范畴，让金融信息安全防范的工作变得更加复杂。

8.2.3 案例：美国“棱镜门”事件

1. “棱镜门”事件回顾

2013 年 6 月，美国前中情局职员爱德华·斯诺顿将两份绝密资料交给英国《卫报》和美国《华盛顿邮报》发表，美国国家安全局有一项代号为“棱镜”的秘密项目要求电信巨头威瑞森公司必须每天上交数百万用户的通话记录，通过进入谷歌、雅虎、微软、苹果、Facebook、美国在线、PalTalk、Skype、YouTube 等九大网络巨头的服务器，监控美国公民的电子邮件、聊天记录、视频、照片等秘密资料，同时，斯诺登称美情报部门 2009 年起开始监控中国内地和香港电脑系统，全世界舆论哗然，中国网络安全堪忧。

棱镜计划(PRISM)是一项由美国国家安全局(NSA)自 2007 年起开始实施的绝密电子监听计划。该计划的正式名号为“US-984XN”。PRISM 计划能够对即时通信和既存资料进行深度的监听。许可的监听对象包括任何在美国以外地区使用参与计划公司服务的客户，或是任何与国外人士通信的美国公民。

受到美国国安局信息监视项目——“棱镜”监控的主要有 10 类信息：电邮、即时消息、视频、照片、存储数据、语音聊天、文件传输、视频会议、登录时间和社交网络资料，具体细节都被政府监控。

通过棱镜项目，国安局甚至可以实时监控一个人正在进行的网络搜索内容。“棱镜计划”项目监视范围很广，主要从美国的网络服务商直接获取相关数据，这些服务商涵盖了互联网行业的多家巨头，包括微软、雅虎、谷歌、Facebook、Pal Talk、YouTube、Skype、AOL 和 Apple(见下图)。

服务商	加入时间
微软	2007-09
雅虎	2008-03
谷歌	2009-01
Facebook	2009-06
Pal Talk	2009-12
YouTube	2010-09
Skype	2011-02
AOL	2011-03
Apple	2012-10

→

监听数据类型
电子邮件
聊天记录
视频、图像
网络存储数据
IP电话
文件传输
视频会议
被监视目标的网络行为
网络社交的具体细节

《华盛顿邮报》披露的“棱镜计划”涉及企业

2. 棱镜门事件所折射出的美国的全球网络空间霸权战略

当今世界，美国作为互联网的发源地和管理大本营，时时刻刻监视着全球各国的一举一动，信息安全就是国家安全。美国严密控制全球互联网，支撑其超级大国的网络空间霸权。自互联网在美国军方的诞生到成熟进化到全球互联网，美国商务部领导的多个非营利机构行使全球互联网管理职责，对互联网的技术标准、管理规范、域名系统、网络地址等进行管理。

2012 年 12 月，在国际电联 178 个国家参与的关于修改《国际电信条约》会议上，由中国、俄罗斯、印度等要求由联合国旗下的国际电信联盟来共同管理全球互联网络，被美国断然拒绝，超级大国的网络空间霸权难以撼动。

1) 美国以占据市场主流地位的高技术公司为先锋，立法贯彻网络空间国家战略

2001 年 10 月 26 日，美国颁布了《美国爱国者法案》。根据法案要求，警察机关有权搜索电话、电子邮件通讯、医疗、财务和其他种类的记录，特别是去掉了对美国本土情报单位的法律约束限制。

按照法案要求，美国“八大金刚”(思科、IBM、Google、高通、英特尔、苹果、Oracle、微软)都或主动或被动地向美当局交付信息。谷歌公开承认已根据《美国爱国者法案》规定，把欧洲资料中心的信息交给了美国情报机构。微软也公开承认美国依法获取欧盟云端资料，毫无悬念，任何一家美国公司，不论是谷歌、微软还是思科，都作为美国的急先锋，在市场经济的合法外衣下，忠实执行着美国网络空间霸权的国家战略。

2) 美国实施全面持续的网络监控计划

“棱镜”(PRISM)项目只是美国政府秘密监控系统的“冰山一角”，仅美国国家安全局(NSA)就实施了 4 项监控项目，并专设了一个 1000 人的情报收集部门“定制入口行动办公室”(TAO)。

4 项监控项目分别为“主干道”(MAINWAY)、“码头”(MARINA)、“核子”(NUCLEON)和“棱镜”(PRISM)项目。“主干道”和“核子”项目负责电信网的基础数据和通话内容的监控，“码头”和“棱镜”项目负责互联网基础数据和通信内容的监控，4 大秘密监视项目帮助美国政府对全球通信进行了有效监控。

(1) “主干道”项目。为美国国家安全局，监视电信网上数以亿兆计的“元数据”，即通话的时间、地点、设备、参与者等，进行存储和分析。美国国安局 2009 年花费 1.46 亿美元购买硬盘等设备，用来存储“主干道”监视项目的元数据。

(2) “码头”项目。为美国国家安全局，监视互联网上数以亿兆计的“元数据”，即通信的时间、地点、设备、参与者等，进行存储和分析。

(3) “核子”项目。为美国国家安全局，专门截获电信网上的电话通话内容。从 2002 年开始，美国 4 大电信运营商 Verizon、AT&T、T-Mobile 和 Sprint 就开始“自愿”与美国国家安全局合作。

(4) “棱镜”项目。为美国国家安全局和联邦调查局，负责截取互联网通信内容。“棱镜”接入谷歌、雅虎、微软等 9 家大型跨国 IT 企业的服务器，截取互联网内容，“定制入口行动办公室”(TAO)，是美国国家安全局下设部门，一直从事侵入中国境内电脑和通信系统的网络攻击，借此获取有关中国的有价值情报。“定制入口行动办公室”

1997 年成立，专门从事秘密侵入外国目标电脑和通信系统，破解密码和安全防火墙，获取和复制目标信息。“定制入口行动办公室”旗下的军事和民间“黑客”、情报分析师、目标定位专家、计算机硬件和软件设计师以及电子工程师总数超过 1000 名，是国家安全局最大、也是最重要的部门。

3) 美国 IT 企业在监控计划中的关键作用

美国 IT 企业在针对网络的监控计划中起到关键作用。微软、雅虎、谷歌、Facebook、PalTalk、美国在线、Skype、YouTube、苹果 9 家大型跨国 IT 企业在 PRISM 计划中占有至关重要的地位。9 家 IT 企业的积极主动配合，使得美国国家安全局可以接触到大量个人聊天日志、存储的数据、语音通信、文件传输、个人社交网络数据。同时，思科等基础设施厂商也参与到棱镜计划，美国国家安全局通过思科路由器监控世界各国网络和电脑，思科的通信设备已分布在全球各大洲各个角落。

美国 9 家 IT 企业先后加入棱镜计划，2007 年 9 月微软公司率先加入棱镜计划，2008 年 3 月雅虎加入，2009 年 1 月谷歌加入，同年 6 月雅虎加入，同年 12 月 PalTalk 加入，2010 年 9 月 YouTube 加入，2011 年 2 月 Skype 加入，同年 3 月 AOL 加入，苹果公司 2012 年 10 月加入棱镜计划。

英特尔旗下信息安全公司 McAfee 就常与 NSA、FBI 和 CIA 合作。McAfee 被视为有价值的合作伙伴，因为该公司能通观恶意互联网流量的情况，包括外国势力的间谍活动。一些黑客利用合法服务器从事黑客活动，而 McAfee 防火墙能收集到这些黑客的信息。此外，McAfee 的数据还能表明一些网络攻击源自哪里。McAfee 同时也了解全球的信息网络架构，这对情报部门来说很有意义。

美国电信运营商在针对电信网的监控计划中起到关键作用。Verizon、AT&T、T-Mobile 和 Sprint 等 4 家大型运营商为美国国家安全局提供了接入国内和国际通信网的“后门”通道，方便 NSA 通过对电信网络进行监听的方式，收集了大量电信数据和很多的交谈信息，在美国监控计划中扮演了重要角色。美国国内最大的电信运营商 Verizon 公司就是“主干道”监控项目的一个原始情报信息提供者，NSA 通过 Verizon 收集数百万美国客户的电话记录，包括美国国内的电话和由国内打往外国的电话。

3. “棱镜门”事件折射出的美国信息战略

1) 美国信息监控计划是一个包含政府、企业和舆论的三位一体长期战略

在美国信息监控计划中，政府、IT 企业和社会团体分别扮演了不同角色，相互配合，默契互动，共同推动网络监控计划实施，保证美国国家安全。一是美国政府部门，如美国国家安全局、国防部和联邦调查局等积极开展监控计划的组织、计划和评估工作；二是美国 IT 企业是美国监控计划的具体实施机构和重要支撑部门，是海量信息和数据的来源，是监控计划的实施主体；三是政府官员、议员、权威专家和非营利组织以保护国家安全为由，对美国监控计划进行声援、游说和宣传。

2) 美国 IT 企业已经成为美国网络战的主力军

网络战呈现出军民融合的趋势，看似平静的和平时代，美国已经通过实施各种计划，发动了网络战争，IT 企业已成为网络战主力军。美国通信巨头思科参与了中国几乎所有大

型网络项目的建设，涉及政府、海关、邮政、金融、铁路、民航、医疗、军警等要害部门的网络建设，以及中国电信、中国联通等电信运营商的网络基础建设，然而思科却是美国政府和军方的通信设备和网络技术设备主力供应商。微软在中国乃至全球都是占有绝对垄断地位的厂商，其 Windows 系列操作系统在我国市场占有率超过九成，其 Windows Phone 手机的操作系统也在我国呈现快速发展趋势。微软在公开发布补丁修复漏洞之前，就会向情报部门提供这些漏洞信息，这些信息可用于保护政府计算机，并入侵恐怖分子或敌对方的计算机。

3) 美国 IT 企业通过深度参与我国信息化建设全面威胁网络空间安全

我国的网络空间安全在以思科为代表的美国大型 IT 企业面前形同虚设，在我国绝大多数核心领域，美国大型 IT 企业都占据了庞大的市场份额。思科的业务已经渗透到国内几大领域的核心企业。中国骨干网络几乎被思科产品全面占据，中国电信 163 和中国联通 169 承担了中国互联网 80%以上的流量，思科占据了中国电信 163 骨干网络约 73%的份额，把持了 163 骨干网所有的超级核心节点和绝大部分普通核心节点，思科占据了中国联通 169 骨干网约 81%的份额。

4. 美国 IT 公司对我国各行业的垄断控制

美国 IT 企业已经在我国骨干网络的基础设备、服务器、个人电脑、手机终端、个人软件系统等行业领域中占据绝大多数的市场份额，其中的大多数处于垄断地位，控制着我国大部分网络和信息系统。

1) 思科在我国市场份额巨大

思科不仅在中国的市场占有率奇高，而且几乎涵盖了我国大部分至关重要的领域。经过 19 年在中国的发展，思科的客户已经遍布了国内几大领域的核心企业，其中包括中国国家金融数据通信骨干网、中国电信、中国联通、中石化、中国人民银行、北京市政府等众多央企及政府部门。

Internet 骨干网络是公众因特网的核心，所有的数据都要经过骨干网进行转发，骨干网络的安全性是电信行业的重中之重。而思科产品占据了中国电信 163 和中联通 169 超过 70%的份额，把持了几乎所有的超级核心节点和绝大部分普通核心节点。除电信行业外，思科在金融行业、政府机构、铁路系统、民航的空中管制骨干网络、电视台及传媒行业都占据了足以形成垄断的份额。

目前思科在中国拥有员工超过 4000 人，分别从事销售、客户支持和服务、研发、业务流程运营和 IT 服务外包、思科融资及制造等工作。思科在中国设立了 12 个业务分支机构，并在上海建立了一个大型研发中心。思科的扩张仍在继续，专家指出：“思科把持着中国经济的神经中枢。有冲突出现时，中国没有丝毫的抵抗能力。”

2) IBM 为我国服务器市场龙头

IBM 在中国服务器市场的占有率为 19.3%，处于第一的位置，其次为戴尔、惠普，几家联合起来占有近八成以上的市场占有率。而联想等仅有不到 10%的市场占有率。IBM 在中国地区的业务目前已深入到服务器、PC、软件、笔记本等多个 IT 领域内，并具有相当的影响和规模。

据国家工商总局公平交易局调查资料显示，在采用英特尔处理器的服务器领域，IBM 市场占有率 19.3%。

3) 英特尔 PC 微处理器市场占有率高

英特尔从来没有公布过在中国的确切销售数字。2012 年统计数据中显示：高居 IT 产业链最上游的英特尔在全球 PC 微处理器市场上的占有率已经扩大到接近 80%。而占总数近 1/3 的最终产品输出到中国。尽管在移动时代英特尔在全球芯片出货量比例有所萎缩，但是在中国 PC 市场，英特尔依旧占有绝对领先的地位。

4) 谷歌安卓系统在我国市场占有率超过八成

对于目前火热的智能手机来说，在中国市场，安卓远远地甩开了苹果 iOS 操作系统以及微软 Windows Phone 以及 Blackberry 10 等，目前在中国市场占有率超过了 8 成。而对于目前的智能手机来说，其安全性以及隐私性的高要求甚至超过了传统的 PC。

5) 微软垄断我国操作系统市场

微软在中国乃至全球都是占有绝对垄断地位的厂商，其 Windows 视窗操作系统，自 Windows 95 以来，几乎垄断了所有的 PC 操作系统。据不完全统计，目前 Windows 7 以及 Windows XP 等市场占有率超过九成。尽管 Windows Phone 手机的操作系统在中国市场占有率不高，但是诺基亚也在中国市场主推其装载了 Windows Phone 操作系统的智能手机。

6) 苹果逐步扩大影响

根据苹果 2012 年财报显示，亚太地区的收入有 2/3 来自中国，上一财季销售总额为 57 亿美元，相比去年同期增长了 48%。苹果 CEO Tim Cook 指出，到上一财季结束，苹果在中国的总收入为 124 亿美元，而 2011 年一年的收入只有 133 亿美元，而这些数字仍然在以难以置信的速度增长。目前 iPad 在中国的市场中占有绝对的领先地位，而苹果的 iPhone 手机也在中国市场中有着很高的占有率，而苹果的笔记本电脑等在中国市场的销售额也在不断增加。

7) 甲骨文垄断我国重要行业数据库市场

甲骨文 1989 年正式进入中国，建立北京首家办事处。目前，甲骨文中国已拥有 2.5 万个客户，4500 名员工，以及 4 个研发中心。经过 20 余年的发展，目前在中国市场上的甲骨文已经控制了 90%的数据库市场。

8) 高通引领移动互联网时代

移动互联网时代的发展造就了高通。高通目前在手机平板等移动设备中占有了相当大的优势地位。根据 iSuppli 调查，高通 2007 年登上全球手机芯片龙头地位后，2012 年市场占有率进一步攀高至 31%，连续 5 年蝉联全球手机芯片龙头。目前在国内的知名厂商中，小米、联想、酷派等大多采用了高通的 CPU。

@ 8.3 大数据金融信息安全风险

8.3.1 大数据金融信息安全风险的类型

在大数据时代，企业金融信息安全面临的风险主要有法律风险、市场风险、技术风

险、操作风险、道德风险等方面。这些风险与大数据技术的发展相依相成，有些风险是大数据与生俱来的固有风险，如物理环境风险和技术风险等；有些风险受大数据技术的外部环境所影响，如法律风险等，有些风险伴随着社会进步慢慢将会得到有效控制，如信息泄密风险等。

1. 法律风险

法律风险是企业在经营过程中由于故意或过失违反法律义务或约定义务可能承担的责任和损失。

法律风险的表现形式如下。

(1) 金融合约不能受到法律应予的保护而无法履行或金融合约条款不周密。

(2) 法律法规跟不上金融创新的步伐，使创新金融交易的合法性难以保证，交易一方或双方可能因找不到相应的法律保护而遭受损失。

(3) 形形色色的各种犯罪及不道德行为对金融资产安全构成威胁。

(4) 经济主体在金融活动中如果违反法律法规，将会受到法律的制裁。

在大数据时代，由于相关法律法规建设尚不健全，存在很多监管漏洞，企业在金融信息安全方面面临着来自法律方面的风险，简单而言，可以表现为以下两个方面。

(1) 大数据产业文化背景带来法律风险。

国内的大数据产业将与欧美完全不同，国外讲究个人隐私，有严格的反隐私法的规定。而东亚文化圈对上网“隐私”容忍度很高，相关法律机制也不健全，也给了一些大数据公司和互联网用数据牟利带来了“空间”。这跟互联网行业早期发展与国内知识产权相对宽松氛围相关，整个行业法律意识相对淡薄，民众版权意识薄弱，知识产权付费使用的意识不强。但是在互联网行业已经相当成熟的今天，法律不健全给企业带来的大数据金融信息安全风险显然已经不容小觑。

(2) 大数据产业的监管漏洞带来法律风险。

大数据是把双刃剑，公民的数据信息必须得到依法监管，一旦出现行业性数据安全泄密事件，将会让相关新行业陷入危机之中。例如，智能家居数据泄密将会造成人身财产安全隐患。比如在 3 月 10 日曝出一起某互联网公司员工人盗取 50 亿条公民数据的信息。这是大数据崛起前最大的绊脚石，也从侧面证明了大数据产业所处的原始混乱状态。在这种混乱状态下，如果行业监管不能得到及时有效跟进，将会给大数据产业的发展带来极大的阻碍，给金融系统带来法律风险和极大的安全隐患。

2. 物理环境风险

物理环境风险是指企业利用大数据技术进行分析所依托的信息系统设施面临的物理环境遭到外部因素影响而给金融信息安全带来的风险。

这些外部因素包括基础设备故障、信息系统故障等。

(1) 基础设备故障给金融信息安全带来的风险。

大数据分析依托的信息系统基础设施包括支撑业务应用系统的网络(局域网、广域网、互联网、专线网、无线网)、硬件(服务器、主机、应用终端、共享设备)和物理环境，它们是组织业务赖以生存的基础(如电力、Web 服务器、数据库服务器等)，一旦出现故障或中

断，它所承载的应用也会出现问题或停顿。基础设备风险要求组织在应用层、网络层、链路层和物理层面进行综合防御。

(2) 信息系统故障给金融信息安全带来的风险。

企业的计算机操作系统和应用软件在组织业务交流的运行过程中，需要一个十分安全稳定的内部环境。来自这些系统和应用软件的问题和缺陷会对系统造成影响，特别是在多个应用系统互联时，影响会涉及整个组织的多个系统，甚至会导致整个公司或网站瘫痪。信息系统风险要求机构对系统应用在协同、系统维护、版本测试、版本管理、配件管理、系统管理、系统监控等方面具备管理能力。

3. 技术风险

大数据时代金融信息安全面临的技术风险是指数据在获取、挖掘、处理等基本环节因技术处理不当或技术设计不到位而引致的风险。

对数据进行收集、存储、处理、挖掘分析是搜索技术的基本环节。金融信息企业主要相关的大数据技术有：数据采集、数据存储、数据处理、数据挖掘与分析技术等。

金融信息安全所面临的技术风险主要体现在以下几个方面。

(1) 完整性风险。即数据未经授权使用或不完整或不准确而造成的风险。这种风险通常与用户界面的设计、数据处理程序、灾害恢复程序、数据控制机制及信息安全机制等有关。

(2) 存取风险。即系统、数据或信息存取不当而导致的风险。在互联网和大数据日益普及的今天，存取风险是企业面临的主要威胁之一。存取风险主要与业务程序的确立、应用系统的安全、数据管理控制、数据处理环境、网络安全、计算机和通信设备状况等有关。

(3) 获得性风险。即影响数据或信息的可获得性的风险。主要与数据处理过程的动态监控、数据恢复技术、备份和应急计划等有关。

(4) 体系结构风险。即信息技术体系结构规划不合理或未能与业务结构实现调配所带来的风险。主要与信息技术组织的健全、信息安全文化的培育、信息技术资源配置、信息安全系统的设计和运行、计算机和网络操作环境、数据管理的内在统一性等有关。

(5) 其他相关风险。即其他影响企业业务活动的技术性风险。主要与信息技术对业务目标的支持、业务流程周期、存货预警系统、业务中断、产品信息反馈系统、业务的流动性管理等有关。

4. 信息泄露风险

大数据时代金融信息安全面临的泄密风险是指数据在获取、存储、传输、分析和使用等过程中发生信息泄露从而给信息相关者带来安全隐患的风险。

信息泄密方式主要有 3 种情况：黑客入侵，用户信息未加密；企业内部员工窃密；服务外包人员窃密。其中，企业员工内部泄密对企业的损害程度和其发生的频度远远高于其他外部攻击窃密，更是防范重点。信息作为组织信息技术系统装载的业务数据，是一种具有非常重要价值的资产。与实物资产相比，信息非常分散，并且容易被复制，信息是组织业务流程中最重要的数据，如客户资料、产品设计等，如果不能得到正确的识别、评估、

保存和管理，就可能面临被窃取、损毁和丢失的风险，这不仅会对依托于这些关键信息的核心业务造成严重破坏，还会对组织的信誉和声望造成巨大的损害，甚至会摧毁整个组织。

近年来，国内网络犯罪案件呈现逐年上升的态势，其中涉及金融业特别是银行信息安全方面的犯罪也不在少数。例如，2014 年 2 月支付宝员工在信息系统的后台下载了大量客户信息有偿出售给其他电商公司；2016 年相继发生的携程信用卡信息泄露、小米社区用户信息泄露等事件中，出现了大量用户信息数据被盗，导致用户网络银行账户被入侵事件等。上述事件严重影响了金融消费者的合法权益，也充分暴露出在网络信息安全领域有较大隐患，不容小觑。

8.3.2 大数据金融信息安全风险的特征

大数据技术的不断发展为金融市场风险监控提供了有效的技术支撑，因而在大数据时代，金融信息安全的风险有着比传统金融信息安全风险更为鲜明的特征。在大数据时代，金融信息安全风险具有扩散性强、影响面广和风险评估难的特点。

1. 扩散性强

由于大数据具有 Velocity(获取及处理速度极快)的特点，在大数据时代，数据的获取是随时随地进行的，与此同时，数据的处理也是飞速的。在大数据的处理过程中，如果某个细微的环节出现错误，这种错误将会以极快的速度蔓延开，扩散能力极强。这是大数据技术与传统海量数据处理的重要区别之一。

大数据时代下金融信息安全风险扩散性强主要体现在以下几个方面。

(1) 大数据技术使得金融机构获取海量数据的过程变得简单和便捷，数据的获取随时随地都在进行。我们在浏览网页时的任何停留都能够迅速被大数据技术捕捉并记录在数据库中。如果有黑客等恶意制造大量虚假数据，这些可以制造的数据将会迅速传播到各数据分析中心，这种虚假数据的传播将会带来极大的金融信息风险，而且这种风险将会以极快的速度扩散。

(2) 在大数据技术的应用下，金融机构处理交易数据和客户数据等的速度和量有了质的提升，机构运行效率提升。如果前面获取的数据存在问题，数据处理时将会得出大量错误的结论。由于大数据处理关注相关关系而不是因果关系，在处理数据时将很难发现有意而为之的数据错误，这种问题带来的风险将会以极快的速度传播到整个数据传输通道。

(3) 随着互联网的普及和大数据的发展，人们获取信息变得更加容易，沟通方式变得更加便捷，消费与购物方式也摆脱了物理形态，通过线上支付，几秒钟就能实现商品交易。若消费者信息被黑客恶意盗用，这种大体量的数据风险将会随着便捷的交易媒介迅速扩散，严重危害到金融信息安全。

2. 影响面广

金融领域对信息变化的反应极为敏感。由于大数据具有体量大、传播速度快等特征，金融市场上一些很细微的操作能被迅速放大并广泛传播，产生“蝴蝶效应”，可能会对资

本市场产生很大的冲击，影响面极为广泛。大数据时代下金融信息安全风险影响面广主要体现在以下两个方面。

(1) 在金融领域，数据与信息的传递速度特别快，金融市场对外界信息的反应程度极大。由于金融全球化，一国金融市场上极小的变动都可能会对全球金融市场上产生重大影响。在大数据时代，这样的影响尤其显著。金融市场具有极强的外部性，容易受外界信息的干扰。

(2) 大数据技术给金融机构带来技术革命，目前许多机构分析数据都依赖于大数据技术。依靠大数据技术，金融机构能在极短时间内对金融市场上的信号做出反应。大数据技术处理数据体量大，速度快，从发现错误到形成实质性损失之间的时间极短，加上金融市场本身固有的脆弱性，使得大数据时代，金融信息安全风险影响力被快速放大。

3. 风险评估难

从金融信息安全的角度来讲，风险评估是对金融信息资产所面临的威胁、存在的弱点、造成的影响，以及三者综合作用所带来风险的可能性的评估。风险评估的主要任务包括：识别评估对象面临的各种风险；评估风险概率和可能带来的负面影响；确定组织承受风险的能力；确定风险消减和控制的优先等级；推荐风险消减对策。信息技术软硬件漏洞是全球各类信息安全问题的主要源头之一，对大数据技术带来的金融信息安全风险评估首先在技术上具有很大难度。另外，就目前而言，并没有一套完善的基于大数据技术带来的金融信息安全风险评估模型。

大数据时代下金融信息安全风险评估难主要体现在以下几个方面。

(1) 从风险揭示层面出发，关于大数据与金融信息安全的相关法律尚不明确，存在很多监管漏洞。就目前而言，大数据技术是一项前沿技术，大数据金融与其他领域的概念可能会发生重叠，导致风险揭示不清晰，风险披露不明朗。依托互联网，大数据的监管更加困难，各国目前也正在积极出台关于大数据金融的监管条例。

(2) 从风险评估步骤层面出发，风险评估包括风险辨识、风险分析、风险评价 3 个步骤。风险辨识是指查找企业各业务单元、各项重要经营活动及其重要业务流程中有无风险，有哪些风险。风险分析是对辨识出的风险及其特征进行明确的定义描述，分析和描述风险发生可能性的高低、风险发生的条件。风险评价是评估风险对企业实现目标的影响程度、风险的价值等。在大数据技术广泛运用的金融机构，从信息采集到数据分析再到生成分析结果，大数据技术的应用贯穿风险评估的每个步骤。

(3) 从风险评估过程层面出发，在风险评估过程中，有几个关键的问题需要考虑。①要确定保护的对象(或者资产)是什么？它的直接和间接价值如何？②资产面临哪些潜在威胁？导致威胁的问题所在？威胁发生的可能性有多大？③资产中存在哪些弱点可能会被威胁所利用？利用的容易程度又如何？④一旦威胁事件发生，组织会遭受怎样的损失或者面临怎样的负面影响？⑤组织应该采取怎样的安全措施才能将风险带来的损失降低到最低程度？解决以上问题的过程，就是风险评估的过程。在大数据时代，这种金融信息安全风险往往十分隐蔽，上述在风险评估过程中的问题很难得到完全的解决。

在大数据时代，金融市场自动化交易发展迅速，利用强大的计算机处理能力，根据交

易模型发出算法指令，具有单笔报单小、报单总笔数高、时间间隔短、报单撤单比高等特点。自动化交易提高了市场流动性和价值发现效率，但也带来一系列风险，且由于交易量庞大，交易时间迅速，交易范围广，所带来的金融信息安全风险影响迅速扩大。在美国期货交易所中，自动化交易成交量占总交易量的一半以上。由于自动化交易普遍采用止损策略，当市场出现大幅波动时，会自动触发一系列相关金融产品的连锁交易，从而引发市场多米诺效应。在大数据时代，由于金融信息安全风险的扩散性强、影响面广、风险评估难的特点，这种高频交易很有可能会迅速导致金融市场全线崩盘，引发资本市场剧烈波动。

在美国关于大数据与自动化交易最著名的案例就是 2010 年 5 月 6 日发生的“闪电崩盘”事件。由于一家交易公司电脑发出错误指令，导致大量自动化交易自动止损，道琼斯工业指数在 30 分钟内狂挫千点，市值损失上万亿美元。2013 年 4 月 3 日，黑客劫持美联社的推特账号，发布了美国白宫发生爆炸、总统奥巴马受伤的假消息，金融市场瞬间出现恐慌性抛售，道琼斯工业指数在 3 分钟内下跌超过 140 点，市值损失近 1400 亿美元。2013 年 8 月 6 日，光大证券由于自动化交易平台缺陷，发送错误指令导致上证指数在 26 秒内狂涨 100 点，造成国内资本市场剧烈波动。

上述案例均表明，大数据在给金融市场带来前所未有的巨大发展的同时，还会带来金融信息安全风险，在大数据时代，这种风险由于具有扩散性强、影响面广而且风险评估难，给金融市场带来很大的挑战。

8.3.3 国内外金融信息安全事件及事故

1. 信息安全事件

信息安全事件是指识别出发生的系统、服务或网络事件，表明可能违反信息安全策略或防护措施失效，或以前未知的与安全相关的情况。

对于金融信息安全事件，由于金融业多金的本质，长期以来，全球各类非法组织、不法分子不断研究和尝试运用各种先进技术手段，利用金融企业管理和金融信息系统的信息安全缺陷和脆弱性，策划和组织金融犯罪活动，资金损失、信息泄密事件层出不穷。此外，金融业一些内部从业人员，也因为利益的驱使，突破道德底线，从内部窃取数据或越权操纵，导致安全堡垒从内部被攻破，内外安全问题夹击，使金融业信息安全更加危机四伏。

2. 信息安全事故

信息安全事故是指一个或系列非期望的或非预期的信息安全事件，这些信息安全事件可能对业务运营造成严重影响或威胁信息安全。

对于金融信息安全事故，金融信息化建设促进金融业信息化程度高度发达，但由于核心业务和核心数据高度依赖信息系统，系统任一环节的运行故障或操作失误都可能会造成严重事故，关键数据的损失可能会对金融企业和金融行业造成致命打击。而信息系统运维失误、外部因素导致的系统运行连续性事故往往是产生金融信息安全事故的主要来源。

3. 国内外金融信息安全案例

表 8.3 收集了互联网上公布的近 4 年来全球范围发生的金融信息安全事件和事故。

表 8.3 近几年来全球金融信息安全事件和事故

年度	信息安全事件
2016	环球银行金融电信协会(SWIFT)信息系统发生多起网络入侵盗窃事件。2016 年 2 月，网络黑客入侵孟加拉国中央银行窃取 8100 万美元；2016 年 6 月，黑客攻破乌克兰银行核心网络系统，盗取 1000 万美元。几宗案件的作案手法十分雷同，均为黑客通过入侵银行账户系统，植入网络木马程序，盗取转账凭证，并篡改 SWIFT 文件，控制交易流程
2015	美国大型医疗保险商 CareFirst 遭遇专业黑客攻击，约有 110 万医疗保险客户的个人信息遭泄露，包括客户的个人姓名、生日、邮箱地址、医疗保险号码等信息，部分信息被发现遭到非法利用。 美国证券服务商 Scottrade 发生了信息系统数据泄露事故。数百万用户的敏感数据受到影响，受影响的数据库中包含用户的社会安全号码和电子邮件地址。 汇丰银行由于内部控制原因，大量秘密银行账户文件被非法盗取，涉及约 3 万个账户，总计 1200 亿美元资产
2014	英格兰银行大额支付系统(CHAPS)故障导致系统宕机长达 10h，事故当日积压大量交易数据，政府、商业和个人的支付业务受到严重干扰，对英国的经济造成重大影响。 欧洲中央银行(ECB)遭到严重的网络攻击，网络黑客通过其外部网站的数据库，窃取了网站上 1.5 亿注册者的电子邮件和用户的个人信息，包括电子邮件、家庭住址和电话号码在内的部分未加密数据被非法利用。 美国第二大零售商家得宝(Home Depot)公司支付系统遭到网络攻击，近 5600 万张银行卡的信息被盗，比 2013 年美国塔吉特(Target Group)发生的客户银行卡数据被盗事件更加严重。 摩根大通银行 7600 万家庭和 700 万小型企业的相关信息被位于南欧的网络黑客盗取，涉及银行客户的姓名、住址、电话号码和电邮地址等个人信息，与这些用户相关的内部银行信息也遭到泄露
2013	2013 年 6 月 23 日，中国工商银行上海数据中心主机运维失误，造成国内多地的网点柜面、ATM、网银业务出现故障，无法办理业务和提供资金服务，故障时间持续 1h，故障涉及北京、上海、广州、武汉、哈尔滨等多个大中型城市，造成较大影响

【案例】数据泄露

2015 年是数据安全事故频发年，也是数据安全防护技术高速发展的一年。回顾整个 2015 年，产业信息化、数字化、网络化进程加速，“互联网+”已然成为一种不可逆的趋势，互联网、云计算、大数据带来更新式革命，然而新趋势下的数据安全状况变得越发严峻。Verizon 新发布的《2015 数据泄露调查报告》显示，500 强企业中超半数曾遭受过黑客攻击。来自中国的数据安全问题更加触目惊心。福布斯上榜的中国企业中，大多数企业都曾经不同程度遭受过攻击或出现数据泄露，特别是一些掌握大量民众个人信息的通信运营商及金融领域。表 8.4 汇总了 2015 年国内外十大最具影响力的数据泄密事件。

表 8.4　2015 年十大国内外数据泄露时间汇总

序号	企业	曝光时间	泄露原因	泄露结果
1	十大知名连锁酒店	2015.02.11	网站存在漏洞，遭受黑客攻击	千万级的酒店顾客敏感信息泄露，包括姓名、身份证、家庭住址、手机号、信用卡等
2	汇丰银行瑞士分支	2015.02.22	内部违规	大量秘密银行账户文件被曝光，涉及约 3 万个账户，总计约 1200 亿美元资产
3	国内社保系统	2015.04.22	大量高危漏洞	数千万人员的身份证、社保参保信息、财务薪酬、房屋等敏感信息
4	工商银行快捷支付	2015.06	存在严重漏洞	多位北京地区的工行储户存款被盗
5	意大利监控厂商 Hacking Team	2015.07	遭受黑客攻击	400GB 内部数据泄露，并可在互联网公开下载和传播
6	美国婚外情网站 Ashiey Madison	2015.08.18	遭受黑客攻击	3700 万名用户资料泄露
7	英国电信运营商	2015.08.09	遭受黑客攻击	240 万用户个人数据及 9 万名客户加密信用卡数据外泄
8	大麦网	2015.08.27	存在安全漏洞	600 余万用户账户密码泄露，并被售卖与传播
9	国家旅游局	2015.09	存在漏洞	6 套系统沦陷，涉及 6000 万客户、6 万多旅行社账号密码、百万导游信息
10	支付宝实名认证	2016.10	存在漏洞	用户实名认证信息下莫名多出 5 个未知账户

1. 十大知名连锁酒店泄露大量房客开房信息

2 月 11 日，据漏洞盒子白帽子提交的报告显示，知名连锁酒店桔子、锦江之星、速八、布丁，高端酒店万豪(丽思卡尔顿酒店等)、喜达屋(喜来登、艾美酒店等)、洲际(假日酒店等)网站存在高危漏洞——房客开房信息大量泄露，一览无余，黑客可轻松获取到千万级的酒店顾客的订单信息，包括顾客姓名、身份证、手机号、房间号、房型、开房时间、退房时间、家庭住址、信用卡后四位、信用卡截止日期、邮件等大量敏感信息。

2. 汇丰发生史上最大规模银行泄密

2 月 12 日，汇丰银行大量秘密银行账户文件被曝光，显示其瑞士分支帮助富有客户逃税，隐瞒数百万美元资产，提取难以追踪的现金，并向客户提供如何在本国避税的建议等。这些文件覆盖的时间为 2005 年至 2007 年，涉及约 3 万个账户，这些账户总计持有约 1200 亿美元资产，堪称史上最大规模银行泄密。

3. 多省社保信息遭泄露，数千万个人隐私泄密

4 月 22 日消息，近日大量社保系统相关漏洞出现在补天漏洞响应平台，网站信息显示深圳、上海、河北、河南、山西、安徽等省市卫生和社保系统出现大量高危漏洞。数据显示，围绕社保、公务员等信息系统的漏洞超过 30 个，涉及人员数量达数千万，其中包括个人身份证、社保参保信息、财务、薪酬、房屋等敏感信息。

4. 工行快捷支付存漏洞，用户存款消失

6 月，工行快捷支付被曝存在严重漏洞，多位北京地区的工行储户遭遇了存款被盗事件。犯罪分子借助非法途径截获短信验证码，轻而易举地盗窃存款。

5. Hacking Team 被黑，“互联网军火”泄露

7 月初，有“互联网军火库”之称的意大利监控软件厂商 Hacking Team 被黑客攻击，400GB 内部数据泄露。据了解，Hacking Team 掌握的大量漏洞和攻击工具也暴露在这 400GB 数据中。更可怕的是，泄露的数据可以在互联网上公开下载和传播。

6. 婚外情网站 Ashley Madison 遭攻击 3700 万名用户资料泄露

8 月，美国婚外情网站“阿什莉 • 麦迪逊”(Ashley Madison)在全世界拥有 3700 万名注册会员，被称为“婚外情界的谷歌”。不明身份的黑客 18 日在网络上公布了这些会员的详细资料，称此举是为逼停网站。黑客公布的资料显示，会员中包括英国公务员、美国银行家和军人以及联合国维和人员等。

7. 英国 240 万网络用户遭黑客侵袭：加密信用卡数据外泄

8 月 9 日，英国电信运营商 Carphone Warehouse 在黑客入侵事件中，包含加密信用卡数据的约 240 万在线用户的个人信息遭到黑客入侵。这 240 万用户的个人数据包括姓名、地址、出生日期和银行卡细节……都有可能遭到黑客访问，其中多达 9 万名客户的加密信用卡数据可能也遭到黑客入侵。

8. 大麦网 600 多万用户账号密码泄露，数据已被售卖

8 月 27 日消息，乌云漏洞报告平台发布报告显示，线上票务营销平台大麦网被发现存在安全漏洞，600 多万用户账户密码遭到泄露。这些隐私数据甚至已被黑产行业进行售卖与传播。

9. 国家旅游局漏洞致 6 套系统沦陷，涉及全国 6000 万客户

该漏洞于国庆长假前夕被补天漏洞响应平台披露，涉及全国 6000 万客户、6 万多旅行社账号密码、百万导游信息；并且攻击者可利用该漏洞进行审核、拒签等操作。通过该漏洞，安全工作者获取了一则长长的名单，能够直接观看到每位用户的详细行程及个人信息。

10. 支付宝实名认证漏洞

10 月，支付宝实名认证存在漏洞。登录支付宝后无意间打开支付宝实名认证页面，用

户的实名认证信息下多出了 5 个未知账户，而且用户没收到任何形式的确认或是告知信息，不论是短信、邮件或者是登录后的站内信息都没有。

从上述总结的政企数据泄密事件来看，主要的泄密风险除了黑客攻击、木马病毒、钓鱼网站等外部因素，缺乏整套行之有效的安全管理系统、内部员工泄密以及内部管理等内部因素成为引发的数据泄密事件的主要诱因。泄密领域也进一步扩大，掌握大量民众个人信息的金融行业依旧是数据泄露的“重灾区”。在“互联网+”时代，企业面临的安全挑战会越来越严峻。随着大数据、云计算以及移动互联网的高度融合，对数据安全技术提出了更高的要求，泄密事件将呈高发势头。

8.4 我国金融信息安全现状及制约因素

8.4.1 我国金融信息安全现状

1. 国家对金融行业信息安全的重视程度不断提高

从政策方面看，党和国家领导人多次就金融行业信息安全做出重要指示，要求金融业研究和把握又好又快的发展规律，努力提高信息安全保障水平，坚决打击危害金融信息安全的犯罪活动。

从资金支持方面看，多年来国家发展改革委等部门针对金融行业信息安全的实际需要，重点支持金融信息安全产品研发和应用等。专项资金的支持在一定程度上有助于提升金融领域信息安全专业化服务水平。

例如，2013 年 8 月国家发展改革委发布《国家发展改革委办公厅关于组织实施 2013 年国家信息安全专项有关事项的通知》，对金融信息安全领域内的金融领域智能入侵检测产品、面向电子银行的 Web 漏洞扫描产品等予以重点支持。2012 年，国家发改委发布《国家发展改革委办公厅关于组织实施 2012 年金融领域安全 IC 卡和密码应用专项有关事项的通知》，对金融领域安全 IC 卡和密码相关关键产品的产业化予以重点支持。2016 年 11 月 7 日，全国人大常委会表决通过《中华人民共和国网络安全法》，该法将于 2017 年 6 月 1 日起施行。《网络安全法》也必将对金融业的发展产生深远的影响。

2. 初步建立以“一行三会”为主的信息安全组织保障机制

中国人民银行着重健全金融信息安全保障体系，联合公安部、安全部、工业和信息化部、电监会四部委共同制定《金融业信息安全协调工作预案》，发布《网络和信息系统应急预案编制指引》，针对区域性电力和通信中断建立联合预警、快速处置流程，并指导省级区域建立信息安全应急协调机制。

银监会将金融业信息技术风险纳入审慎监管整体框架。以《商业银行信息科技风险管理指引》为核心，建立了针对突发事件、业务连续性、科技外包等的监管制度；实施信息科技现场检查和非现场监管，推荐监管评级；同时建立与公安机关、中国银联、电力、电信、证券等部门以及重要信息系统服务商的安全突发事件应急协调机制，加强情报交流与技术协作，提高信息安全协同保障能力。

证监会制定并采取了“纵深防御，平战结合”的防护策略，建立健全信息安全监管制度的同时，推进行业技术基础设施建设，实现了全行业数据集中备份，在应急响应方面开展信息安全应急演练，不断提高应急处置能力。

保监会在落实国家信息安全等级保护的基础上从多方面加强信息安全监管体系建设。一是早在 2008 年就发布了《保险业信息系统灾难恢复管理指引》，又于 2011 年出台《保险公司信息系统安全管理指引(试行)》，对客户信息安全也出台了相关办法并加强管理，同时还建立保险信息安全风险评估指标体系。二是建立跨部委合作机制，制定应急协调预案，加强安全协调与通报工作。三是开展保险业信息系统安全大检查，查补漏洞，提高应急处置能力。

3. 以密码技术和身份认证为主的安全技术保障能力不断加强

当前基于 PKI 的信息安全产品已经成为保障我国金融行业信息安全的有力武器。

(1) 金融机构利用 PKI 机制可以实现用户身份的鉴别，基于 PKI 技术的数字证书已经成为保障网络金融交易的主要工具。通过 PKI 技术加强身份认证、严格控制登录者的操作权限，实现对操作系统和应用系统严格的授权管理和访问控制机制。

(2) 通过采用服务器证书可以实现对网站的可信性认证，有效防范网站钓鱼等金融诈骗。

(3) 手机短信、动态令牌等安全产品也一定程度上保障了金融交易的安全，并得到广泛应用。

(4) 人民银行还针对 RSA1024 算法破解、数据同步机制促发系统停机，云灾备安全风险、支付空间的漏洞、银行卡交易信息截取等方面的问题开展了研究。

4. 已形成移动支付、信息安全等级保护等方面的系列标准

自 2011 年以来人民银行积极研究规划移动支付标准体系，目前已形成涵盖应用基础、安全保障、设备、支付应用、联网通用 5 大类 35 项标准在内的中国金融移动支付标准规范体系。

信息安全等级保护等方面的系列标准也逐步完善。依据《信息安全等级保护管理办法》，人民银行出台了《中国人民银行关于银行业金融机构信息系统安全等级保护等级的指导意见》，并于 2012 年发布了《金融行业信息系统信息安全等级保护实施指引》《金融行业信息安全等级保护测评服务安全指引》《金融行业信息系统信息安全等级保护测评指南》3 项行业标准，在采用《信息系统信息安全等级保护基本要求》的 590 项基本要求的基础上，补充细化基本要求项 193 项，新增行业特色要求项 269 项，为金融行业开展关键信息系统信息安全等级保护实施工作奠定了坚实基础。

5. 信息安全等级保护工作稳步推进

截至 2012 年年底，全国性银行业金融机构完成了 880 个二级以上信息系统的定级评审。2013 年，人民银行发布了《中国人民银行办公厅关于开展重要信息系统信息安全等级保护测评整改工作的通知》，启动了全行范围的重要信息系统等级保护测评整改工作。测评范围为反洗钱中心、征信中心、清算总中心和金融信息中心的 48 个重要信息系统。

据测评机构统计，通过测评，各单位共发现了 4284 项安全问题，整改完成了 3451 项。通过测评整改，各单位普遍增强了信息安全意识和工作技能，信息系统安全管理水平、安全防护能力得到显著提高，人民银行重要信息系统的测评符合率平均值达到了 90% 以上。

8.4.2　我国金融信息安全的制约因素

1. 金融信息技术对外依赖程度较高

目前国内金融业使用的信息系统和网络设备，大部分来自国外，包括数据存储器、操作系统、数据库、芯片等。由于不掌握核心技术，很难判断设备是否存在开发中有意预留或无意疏忽造成软件陷阱等安全漏洞。另外，新技术在带来金融业务增长的同时本身也会带来风险。在特殊情况下，安全漏洞可能被利用实施入侵，修改或破坏设备程序，或从设备中窃取机密数据和信息。

2. 金融信息安全保护的法律环境缺失

我国现代征信体系建设起步较晚，征信管理立法更加滞后。征信是一项法律性很强的工作，由于对企业和个人信息主体征信涉及公民隐私和企业商业秘密等问题，而我国现有的法律体系中尚无一项法律和法规为征信活动提供直接依据，导致征信机构在信息采集、信息披露等关键环节上无法可依，征信主体权益难以保障，严重影响了我国征信体系的健康发展。

近年来，黑客把攻击银行、证券等金融机构信息作为网络违法犯罪活动的重要目标。基于开放性网络的金融服务一旦发生风险，可能造成客户重要数据丢失，使客户资金处于危险状态。而我国在金融信息安全保护立法方面的缺陷，导致监管手段和措施乏力，金融信息的安全与保护面临巨大的风险和挑战。

3. 金融业信息安全联动机制有待加强

金融信息的安全与保护是一个综合性和复杂性的社会工程，需要多个职能部门加强分工协作，密切沟通配合。而我国在金融信息保护工作中，多方联动、上下齐抓的工作机制没有形成，应急管理体系和职能划分制度不完善，也未能建立有效的评估和审议工作制度，金融信息安全防护工作处于金融机构独立管理和维护状态，势单力薄，一旦发生重大应急性信息安全事件，将对我国金融业整体稳健运行造成很大的冲击和影响。

4. 来自外部的风险威胁增多

除开放式网络操作性风险外，外部金融力量入境也给我国金融信息安全造成了潜在隐患，如：世界四大会计师事务所已控制并试图垄断中国的会计审计业； 三大评级机构在中国积极拓展业务； 国际投行对中资企业境外上市的咨询承销已形成垄断； 国际战略投资者的引进使中资金融机构的投资经营活动等信息呈现“客观外泄”。因此，金融企业加强自身信息安全保障工作，建立完善的安全机制来抵御外来和内在的信息安全威胁就显得尤为必要和紧迫。

8.5 美国金融信息安全保障机制

现代金融业作为知识密集型产业，在目标规划、研发建设、运行维护、监控或退出与信息技术相关的产品、服务传递渠道等方面日益体现出以知识和信息技术为基础的特征。金融作为现代经济的核心，关系着国家安全、经济命脉和社会稳定，而金融信息安全则成为影响政治、经济等国家战略安全的重要因素。

20 世纪中期以来，美国在注重信息优势发展经济的同时，将信息安全纳入国家安全战略范畴，建立起一整套较为完善的信息安全保护和防范机制，美国在金融信息安全保障机制方面的成功做法值得我国学习借鉴。

8.5.1 美国金融信息安全保障机制的特点

一直以来，美国把信息安全问题列为国家安全战略的最重要组成部分，在推进信息技术与金融业务融合发展的同时，对关系金融业命脉的数据信息在经济金融全球化趋势下提出了更高的风险管理要求，建立的信息保障机制从技术基础、信息运营系统、管理模式等方面都体现出了先进的理念和特点。

1. 顶级的信息安全技术基础

美国拥有全球顶级的 IT 企业和人才，IBM、EMC、ORACLE 等公司作为全球数据存储系统的垄断寡头，其高端的数据存储技术为金融业的信息安全奠定了坚实基础，为以金融行业为首的众多行业提供优秀的数据存储服务和 IT 解决方案，客户群遍及全球。

美国 IT 软硬件公司与科技专家高度重视技术革新对金融信息安全的保障作用，不懈追求新技术，目前已拥有 3EB 容量的磁存储技术、存储速度更快的热储技术等高端技术。

美国银行更是不惜花费成本更新系统的硬件和软件，为金融数据信息打造了一个安全的“避风港”。

2. 完善的金融信息系统

美国金融业在利用信息技术推动管理和业务创新的同时，注重加强信息系统安全防护，构筑能适应庞大、集中的金融数据处理与传输要求的安全屏障。

美国的金融业内部、金融业之间、金融业与客户三层信息系统已相当完善，数据备份、加密技术、访问控制、入侵检测、漏洞扫描、防病毒等安全保障措施到位。每个信息系统都建立了标准化的操作规则，既提高了金融机构的管理效率和服务质量，也最大限度地避免了联网带来的病毒感染、黑客攻击、身份假冒等安全威胁。

3. 实现了动态和持续化管理

信息技术的持续更新决定了对风险的识别、管理和控制需要动态及时跟进，这也是实现信息完整性、保密性及可用性的必然要求。

美国的金融信息安全保障体制已形成动态管理模式，如著名的美国 RSA 信息安全公司就提供专门的外部安全服务，帮助金融机构应对各种安全威胁，既包括日常安全监测维护，也包括在发生欺诈、钓鱼攻击、僵尸网络等突发安全事故时提供实时应急保护，将威胁快速阻止并将危害降到最低，并配合相关部门开展对入侵者的事后追踪。

8.5.2 美国金融信息安全保障机制的主要做法

1. 完善金融信息安全政策立法

美国政府采取多角度形式构筑金融信息安全的政策立法。

1966 年《信息自由法》将金融信息列为需要保护的信息之一。1996 年《经济间谍法案》《国家信息基础设施保护法案》等规定未经授权、基于商业目的进入在线的计算机窃取金融信息可判监禁最高为 20 年的重罪。1997 年颁布的《关于信息安全技术及产品对外国政府开放的管理规定》在出口产品方面对加密软件产品、高端技术产品严加管制。美国还运用行政权力保障金融信息安全，如《克林顿政府对关键基础设施保护的政策》，布什在任期内签署《信息时代的关键基础设施保护》，并督促国会通过《联邦信息安全管理法案》等，这些都是为了保护通信、金融、能源等基础设施信息的安全。

2. 推行信息安全产品评估策略

美国政府将信息产品的安全性和可信度作为信息基础建设的重要组成部分，早在 20 世纪 70 年代就开展了信息产品安全性评估的研究，20 世纪 90 年代末设立了专门从事信息产品安全评估的机构——国家信息保障同盟(NIAP)。

NIAP 由国家信息与技术研究所以及国家安全局的专业技术和管理人员组成，代表国家指导和监督信息产品的安全评估工作。其下设的计算机安全事业部负责信息产品脆弱性研究与信息安全技术开发，制定有关信息技术标准、制订测试与评估方法及实施方案。美国对信息安全产品的评估策略实际上是将信息安全技术和产品的发展置于政府的完全监督和控制之下。

3. 立体和层次化的金融信息安全管理体系

在国家层面，美国设立了行政实体“总统关键基础设施保护办公室”作为联邦基础设施安全(包括金融信息安全)保护的最高管理协调机构，通过定期集会，加强关键基础设施安全保护中公共和私营部门间的合作，并在必要的时候向总统提交报告。

在部委层面，除涉及国防、外事、情报、执法等关键领域的保护职能必须主要由联邦政府执行外，针对其余每一个信息基础设施部门，指定一个唯一的联邦部局作为领导机构负责协调美国政府在该领域内的活动。

在机构层面，各联邦机构自己负责机构内的信息安全保障工作，将信息安全管理纳入机构战略和运营规划，并定期向机构主管，众议院、参议院、国会授权的对口委员会以及审计总署提交报告，汇报机构信息安全策略实施情况。

4. 统筹相关职能部门并明确职责

金融信息安全工作牵涉部门多，美国统筹规划和明确职能部门对信息安全工作的职责，确保金融信息安全管理高效统一。

财政部负责银行与金融信息安全的统筹协调；科技政策办公室负责协调安全保护方面的科研工作；管理和预算办公室负责监督联邦政府的计算机安全制度在整个政府部门的实施，并每年对信息安全程序和实践进行有效性测试及评估；中央情报局负责评估其他国家对美国网络和信息系统的威胁； 司法部和联邦调查局则负责对网络和信息犯罪的调查和起诉工作。

8.6 我国金融信息安全建设

8.6.1 完善顶层设计，尽快构建适应我国金融发展需要的金融信息安全保障体系

我国金融行业被美国高技术公司全面渗透，部分产品或服务呈现垄断态势。思科公司自 1997 年进入中国就全面参与金融核心骨干网建设，其网络设备在金融行业广泛使用。思科在华金融业务广泛，客户包括工商银行、农业银行、中国银行、地方银行等，占据主流服务提供商位置。IBM 面向工商银行、建设银行、农业银行，以及一些地方银行等全面提供金融业务系统的建设、规划、咨询、方案、产品、运维等，占据市场核心位置。微软的服务器和终端操作系统占据了垄断地位。在当前的网络安全环境下，国家金融在网络空间中的安全风险不可忽视，应尽快研究制定国家金融行业信息安全的防护体系。

《2006—2020 年国家信息化发展战略》明确提出，“要把信息化作为覆盖现代化建设全局的战略举措”。“十二五”期间，我国的金融业将进一步强化“科技保障业务、科技引领业务”的能力，努力赶超发达国家金融信息安全体系水平，达到 “使用面广，设备先进；功能齐全，服务完善；自动化程度高，安全保密性强”。

构建金融信息安全体系的总体目标是物理安全、网络安全、数据安全、信息内容安全、信息基础设备安全与公共信息安全的总和，最终目标是保障业务持续，促进业务发展，保障信息的机密性、完整性和可用性，以及信息系统主体对于信息资源的控制。金融信息安全体系的构建必须符合国家和金融管理部门有关信息安全的政策、标准、规范、指南和细则。金融信息安全体系主要包括以下 4 个领域：信息安全策略、信息安全管理、信息安全运作和信息安全技术。

8.6.2 尽快制定我国金融行业国产信息技术产品和服务替代战略

我国金融行业信息化建设呈现对国外厂商依赖进一步加深的倾向。鉴于业务连续性和高可靠性等要求，我国金融业信息系统和业务系统大量采用了国外厂商生产的设备和系统，如服务器、小型机、大型机、存储设备、网络设备、芯片以及操作系统、数据库、密码算法、安全通信协议等，覆盖了金融核心业务系统运行和服务的各个环节，成为不可替

代的系统，而且随着信息系统不断升级，越来越依赖于国外厂商，这种趋势正在逐步加深。因此，十分迫切需要金融行业信息技术产品和服务的国产化，以便为我国金融信息安全体系建设提供可靠的技术保障。

8.6.3 尽快制定金融行业自主可控战略实施步骤，推进自主可控国家战略

我国金融行业信息安全战略难以根本上保障，亟须完善自主可控的国家信息安全体系建设。我国网络系统大多依托美国公司的技术、装备和服务，它们对于监控者来说几乎是透明的，更不用说其自身具有的“后门”。目前，常规安全防护措施已经无济于事，必须从根本上改变这种局面，调整国家战略进行信息安全保障。从“棱镜门”事件折射出，我国亟须完善自主可控的国家信息安全体系建设，紧跟国家信息安全等级保护制度，强化基础网络和重要信息系统的等级化保护和监督管理，落实等级保护相关措施。同时，鼓励和扶持民族核心技术及产品创新，运用具有自主知识产权的产品和技术，保障国家基础网络和重要信息系统安全，实现真正的自主可控。

8.6.4 应用大数据进行信息安全分析

应用大数据平台进行信息安全分析，通过对安全日志、应用日志、业务数据、外部数据进行风险关联分析，及时发现来自外部的攻击行为、内部违规行为；通过外部泄露数据与银行客户、交易数据的分析，主动发现针对客户的攻击行为、识别客户的潜在风险，对高危风险进行预警；通过对海量历史数据挖掘分析及智能学习，还原客户、用户历史操作行为，获取风险模型的新型特征，使大数据助力于信息安全。

首先，可建设案件分析实验室，通过对已发生案件数据收集，在大数据平台的实验数据环境下进行模型验证与训练，寻找案件的典型行为特征。系统通过一段时间在客户登录、操作、交易过程中对该特征行为规则的分析跟踪优化，将成熟的风险模型运用在监控系统上。

其次，可建设基于大数据的安全威胁情报监控系统，实现对安全事件的有效预测和自动化实时控制，及时发现安全威胁、识别潜在安全隐患，把握安全风险态势，由被动的安全防御向主动的事前安全防御转变。

在新形势下，搭建自身的安全防护体系、设计安全规划，需要将终端、云端、网端三位一体的综合协同联防的安全防御思路融入其中，充分利用大数据进行威胁情报的数据集中与深度挖掘，才能有效应对大数据时代的各种新型威胁，保护金融业的重要信息资产。

本章总结

- 金融信息安全是指利用信息或者金融数据信息，对金融领域实施的各类安全措施和应对手段。金融信息安全包括数据安全、运行安全、软件安全和物理安全。
- 金融信息安全与一般的信息安全相比有较多的相似性，具有保密性、完整性、可用性、真实性、可追溯性、可靠性保护以及连续性等特点。

- 金融信息安全是国家安全中重要的根本内容之一，金融掌握着国家的经济命脉，没有良好的金融安全保障，就会危及国家的安全稳定。另外金融信息安全更是国家发展战略的重要基石，金融数据信息的破坏和窃取往往会对国家、社会。个人产生巨大的损失。
- 大数据在应用我国信息安全时也存在一定的隐患：①大数据的应用会侵犯客户的个人隐私。②数据监听会威胁国家金融安全。③虚假数据会导致金融市场异常敏感。④国内相关法律的缺失存在风险。⑤金融信息安全技术的发展跟不上层出不穷的互联网应用发展速度。⑥网络安全防控是互联网金融信息安全防范的难点。
- 大数据金融信息安全风险主要包括法律风险、物理环境风险、技术风险、信息泄露风险等，具有扩散性强、影响面广以及风险评估难等特点。
- 我国逐渐加大了对金融信息安全的重视，已经形成以“一行三会”为基础的安全保障机制，相关的信息技术也不断地进步。但我国金融信息技术的对外依赖程度依旧较高、缺乏良好的法律环境、金融业信息安全联动机制不完善以及外部风险威胁增大，因此，金融信息安全依旧任重而道远。

本章作业

1. 国际上没有对信息安全的一致定义，请问你是如何理解信息安全的？
2. 简述金融信息安全的定义及属性特征。
3. 你觉得金融信息安全重要吗？为什么？
4. 众所周知，大数据给金融信息安全带来了机遇的同时，也带来了巨大风险，谈谈你认为如何才能高效率使用大数据。
5. 你认为我国金融信息安全保护在现阶段有哪些制约因素？如何打破这些制约因素？
6. 谈谈如何学习美国的金融信息安全保障机制，来构建中国特色的金融信息安全体系。
7. 叙述你感兴趣的一个金融信息安全事件或事故，并说说你从中得到了什么启示。

参 考 文 献

[1] 佚名. 金融大数据[M]. 上海：上海科学技术出版社，2014.
[2] 王和. 大数据时代保险变革研究[M]. 北京：中国金融出版社，2014.
[3] 李勇，许荣. 大数据金融[M]. 北京：电子工业出版社，2016.
[4] 陈云. 金融大数据[M]. 上海：上海科学技术出版社，2015.
[5] 陈红梅. 互联网信贷风险和大数据[M]. 北京：清华大学出版社，2015.
[6] 许伟，梁循，杨小平. 金融数据挖掘：基于大数据视角的展望[M]. 北京：知识产权出版社，2013.
[7] 陈利强，梁如见，张新宇. 金融大数据：战略规划与实践指南[M]. 北京：电子工业出版社，2015.
[8] 何裕. 基于数据挖掘组合模型的股价预测研究[D]. 四川：西南财经大学，2014.
[9] 张丽. 投资者情绪对股票收益波动影响的实证研究[D]. 沈阳：沈阳理工大学，2014.
[10] 赵顺乾. 证券客户关系管理系统应用研究[D]. 上海：上海复旦大学，2013.
[11] 黄斌. 投资者情绪对股票收益的影响研究[D]. 南昌：江西财经大学，2013.
[12] 胡林林. 基于数据挖掘技术的股价指数分析与预测研究[D]. 四川：西南财经大学，2013.
[13] 朱博雅. 一种基于数据挖掘的量化投资系统的设计与实现[D]. 上海：上海复旦大学，2012.
[14] 张静. 智能选股及股价预测系统研究与开发[D]. 长沙：中南大学，2010.
[15] 唐文慧. 基于数据挖掘技术的股价预测实证分析[D]. 四川：西南财经大学，2009.
[16] 王峰. 数据挖掘在证券公司客户关系管理中的应用[D]. 哈尔滨：哈尔滨工程大学，2008.
[17] 刘静. 基于数据挖掘的证券公司客户细分及其应用研究[D]. 上海：同济大学，2008.
[18] 罗月丰. 基于数据挖掘的证券 CRM 客户细分研究[D]. 北京：中国地质大学，2006.
[19] 马杰. 大数据征信应用于互联网金融风控研究[D]. 北京：对外经济贸易大学，2015.
[20] 甘强. 基于混合算法的 P2P 网贷产品推荐系统的设计与实现[D]. 北京：中国科学院大学(工程管理与信息技术学院)，2015.
[21] 孟欣. 大数据时代互联网金融现状及影响分析[D]. 天津：天津财经大学，2014.
[22] 冯晓龙. 基于用户行为分析的 P2P 流媒体推荐系统研究[D]. 北京：北京交通大学，2013.
[23] 姜俊琳. 大数据时代的征信创新与发展研究[D]. 浙江：浙江大学，2016.
[24] 于欣言. 大数据在互联网金融征信中的应用研究[D]. 北京：首都经济贸易大学，2016.
[25] 胡增圣. 数据挖掘方法与股价预测[D]. 北京：中国科学技术大学，2015.
[26] 刘定平. 突发事件环境下投资者情绪对股票价格波动影响的实证研究[D]. 四川：西南财经大学，2014.
[27] 王永红. 切实加强金融信息安全管理提升金融信息安全保障水平[J]. 中国信息安全，2013(4).
[28] 王裕. 基于云平台的大数据处理流程的关键技术研究[J]. 信息技术，2014(9).
[29] 徐计，王国胤，于洪. 基于粒计算的大数据处理[J]. 计算机学报，2015(8).
[30] 唐文方. 大数据和小数据[J]. 金融博览，2015(10).
[31] 梁威. 大数据在零售业的应用[J]. 信息与电脑，2014(8).
[32] 孙慧. 大数据在医疗行业中的应用与挑战[J]. 解放军医院管理杂志，2015(11).

[33] 程立国，陈健恒，徐永红. 大数据在金融业的应用初探[J]. 中国金融电脑，2013(10).
[34] 陈静，孙中东，林磊明等. 大数据驱动金融业变革[J]. 金融电子化，2013(12).
[35] 李庆莉. 大数据驱动银行业智慧化变革[J]. 中国金融电脑，2015(10).
[36] 王和，鞠松霖. 基于大数据的保险商业模式[J]. 中国金融，2014(15).
[37] 梁立. 大数据时代保险业的发展[J]. 时代金融，2014(7).
[38] 刘新海，丁伟. 大数据征信应用与启示——以美国互联网金融公司 ZestFinance 为例[J]. 清华金融评论，2014(10).
[39] 沙莎. 大数据在金融行业的应用[J]. 中国金融电脑，2014(6).
[40] 白运会. 大数据时代的金融信息安全[J]. 网络安全技术与应用，2014(11).
[41] 孔德超. 大数据征信初探——基于个人征信视角[J]. 现代管理科学，2016(4).
[42] 于晓阳. 互联网+大数据模式下的征信 ——以芝麻信用为例[J]. 北方金融，2016(11).
[43] 刘新海. 传统个人征信机构的大数据征信——以环联为例[J]. 清华金融评论，2015(9).
[44] 刘颖，李强强. 从蚂蚁金服看大数据背景下互联网金融征信的兴起[J]. 河北金融，2016(2).
[45] 陆岷峰，虞鹏飞. 互联网金融背景下商业银行“大数据”战略研究——基于互联网金融在商业银行转型升级中的运用[J]. 经济与管理，2015(3).
[46] 黄昶君，王林. 大数据助推银行零售业务量化经营——大数据时代的零售数据挖掘和利用探索[J]. 海南金融，2014(1).
[47] 雷晨光，陈运娟. 大数据时代下商业银行客户关系管理思维变革[J]. 金融与经济，2015(4).
[48] 张宁. 大数据背景下寿险产品定价与创新[J]. 金融经济，2014(2).
[49] 张宁. 云计算在保险公司信息化中的应用[J]. 数学的实践与认识，2012(27).
[50] 朱星婢，何径沙. 大数据安全现状及其保护对策[J]. 信息安全与通信保密，2014(11).
[51] 尹会岩. 保险行业应用大数据的路径分析[J]. 上海保险，2014(12).
[52] 何建雄. 建立金融安全预警系统——指标框架与运作机制[J]. 金融研究，2001(1).
[53] 陈希，李迪安，高星，陈帅，谢邦昌. 数据挖掘技术在保险客户理赔分析中的应用[J]. 统计与决策，2010(4).
[54] 宋华，陈思洁. 供应链金融的演进与互联网供应链金融：一个理论框架[J]. 中国人民大学学报，2016(5).
[55] 何飞，张兵. 互联网金融的发展：大数据驱动与模式衍变[J]. 财经科学，2016(6).
[56] 侯富强. 大数据时代个人信息保护问题与法律对策[J]. 西南民族大学学报(人文社科版)，2015(6).
[57] 史金召，郭菊娥. 互联网视角下的供应链金融模式发展与国内实践研究[J]. 西安交通大学学报(社会科学版)，2015(4).
[58] 辜明安，王彦. 大数据时代金融机构的安全保障义务与金融数据的资源配置[J]. 社会科学研究，2016(3).
[59] 邵建利，宋宁，张滟. 电子商务中第三方支付平台欺诈风险识别研究[J]. 商业研究，2014(11).
[60] 魏强. 大数据征信在互联网金融中的应用分析[J]. 金融经济，2015(8).
[61] 董小君. 美国金融预警制度及启示[J]. 国际金融研究，2004(4).
[62] 周路菡. 棱镜下的大数据恐慌[J]. 新经济导刊，2013(9).
[63] 陈明奇，姜禾，张娟，廖方宇. 大数据时代的美国信息网络安全新战略分析[J]. 信息网络安全，2012(8).

[64] 彭宇，庞景月，刘大同等. 大数据：内涵、技术体系与展望[J]. 电子测量与仪器学报，2015(4).
[65] 赵森林. 大数据的内涵及价值分析[J]. 中共马鞍山市委党校学报，2015(3).
[66] 李芬，朱志祥，刘盛辉. 大数据发展现状及面临的问题[J]. 西安邮电大学学报，2013(5).
[67] 陈静，孙中东，林磊明等. 大数据驱动金融业变革[J]. 金融电子化，2013(12).